DISSERTATIONES
INFORMATICAE
UNIVERSITATIS
TARTUENSIS
25

KAIDO LEPIK

Inferring causality between transcriptome and complex traits

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

25

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS
25

KAIDO LEPIK

Inferring causality between transcriptome and complex traits

UNIVERSITY OF TARTU
Press

Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Estonia.

Dissertation has been accepted for the commencement of the degree of Doctor of Philosophy (PhD) in informatics on 15 December 2020 by the Council of the Institute of Computer Science, University of Tartu.

Supervisors

Dr. Hedi Peterson
Institute of Computer Science, University of Tartu
Tartu, Estonia

Prof. Jaak Vilo
Institute of Computer Science, University of Tartu
Tartu, Estonia

Opponents

Prof. Jack Bowden
University of Exeter Medical School, University of Exeter
Exeter, United Kingdom

Prof. Samuli Ripatti
Institute for Molecular Medicine Finland (FIMM), HiLIFE,
and Faculty of Medicine, University of Helsinki, Finland
Broad Institute of MIT and Harvard, Cambridge, MA, USA

The public defense will take place on 4 May 2021 at 14:15 at Narva mnt 18 and it will be accessible online via Zoom.

The publication of this dissertation was financed by the Institute of Computer Science, University of Tartu.

ISSN 2613-5906
ISBN 978-9949-03-574-8 (print)
ISBN 978-9949-03-575-5 (PDF)

University of Tartu Press
http://www.tyk.ee/

But I don't want to go among mad people, Alice remarked.

Oh, you can't help that, said the Cat: *we're all mad here. I'm mad. You're mad.*

How do you know I'm mad? said Alice.

You must be, said the Cat, *or you wouldn't have come here.*

Lewis Carroll *(Alice in Wonderland)*

ABSTRACT

All things are difficult before
they are easy

Thomas Fuller

In order to pursue a long and healthy life, efficient intervention strategies are required to tackle common age-related inflammatory diseases (e.g. cardiovascular diseases). Though lifestyle factors play an important role in keeping healthy, disease traits are fundamentally governed by genetic processes. Modern medicine exploits this knowledge through development of drugs that manipulate disease progression. As drugs work by targeting gene products, causal genes need to be identified. The main quality standard for making sure that drugs work as intended is provided by randomized controlled trials. Unfortunately, these are usually very time consuming and expensive to carry out, and in some cases might come with additional considerations like ethics and feasibility. Faster and more robust approaches could help deliver a breakthrough in tackling disease. Statistical methods that enable to identify causal gene-trait relationships based on observational data are very promising in this regard.

In this dissertation, we prioritize causal genes for complex disease-relevant traits using statistical methods. This is a rapidly evolving field in statistical genetics which has recently been propelled by population sampling in very large scale by national biobanks. Much of the relevant mathematical theory for causal inference is yet scattered across different disciplines of science: traditional statistics, econometrics, genetics, and the theory of causal inference. A major contribution of this dissertation—next to research contributions published in peer-reviewed journals—is to harmonize this theory under the same causal framework for the purposes of gene prioritization. As such, we spend a considerable amount of effort on introducing concepts and building mathematical structure for causal inference from the ground up. The methodology that we cover is then expanded and utilized in applied settings on human data.

Robust causal inference necessitates sample sizes in the thousands. As part of our research contributions, we worked out a likelihood-based causal model selection approach that allows to prioritize functional genes in smaller samples ($n \approx 500$). We applied our method in the analysis of an inflammatory biomarker C-reactive protein and identified its causal regulatory effect on *CD59* expression—and thus a potential protective effect on healthy cells during immune response. To widen our reach, we developed an algorithm to identify causal relationships for arbitrary phenotypic traits among the entire set of genes. Applied on 43 human traits, we uncovered thousands of novel gene-trait causal links. In a careful analysis of a particularly troublesome genomic region (16p11.2) with many potentially disease-relevant genes, we could pinpoint causal genes for sexual development (*ASPHD1* and *KCTD13*). We further set out to explore how much of sex differ-

ences in traits can be attributed to sex differences in gene expression regulation; however, power analyses show that larger sample sizes are required to provide a definite answer to this question. Finally, we show that gene expression-trait correlations tend not to match up with causal gene-trait relationships, providing additional proof that correlations should not be used for causal gene prioritization. We theorize that instead of trait-influencing genes, trait-affected genes could share a higher overlap with trait-correlated genes.

Statistical methods invariably rely on assumptions and these can be hard to validate, especially in the context of causality. In our work, we add considerable conviction to the findings by using multiple orthogonal approaches together with lab experiments to study the same phenomena, attempting causal reasoning only after triangulation of evidence. This is crucial to minimize the risk of false positive results stemming from unverifiable method assumptions that do not hold in practice. The final verdict of causality needs to be delivered in the lab or by randomized controlled trials. Nevertheless, methods of causal inference hold a lot of potential in providing promising hypotheses that could be prioritized in further studies. The efficiency of computational analyses is key to advancing disease intervention strategies in unprecedented scale and pace.

CONTENTS

List of original publications **13**

1. Introduction **15**
1.1. Applying genotyping in medical genetics 15
1.2. Identifying intervention candidates for complex diseases 16
1.2.1. Identifying trait-associated causal genes 17
1.3. Aims of the dissertation 18

2. Fundamentals of biological principles **20**
2.1. Cells maintain life using proteins 20
2.1.1. Governed by the central dogma of molecular biology . . . 20
2.1.2. Proteins' involvement in disease through gene expression . 21
2.2. The basics of genetics 23
2.2.1. DNA inheritance patterns 23
2.2.2. Genetic architecture of complex traits 24

3. Fundamentals of statistical genetics **26**
3.1. Genome wide association study 26
3.1.1. Ordinary least squares estimator 27
3.1.2. Meta-analysis 28
3.2. Summary statistics 30
3.2.1. Standardization 31
3.2.2. Binary outcome 31

4. Association based gene prioritization **33**
4.1. Fine-mapping 33
4.1.1. Stepwise conditional analysis 34
4.1.2. Bayesian fine-mapping 36
4.2. Colocalization 37
4.2.1. Bayesian colocalization 37
4.2.2. Non-Bayesian colocalization and ties to causality 38
4.3. Transcriptome-wide association studies 39
4.3.1. TWAS for implicating causal genes 41

5. Causal inference **42**
5.1. Causal relationships 42
5.1.1. Directed acyclic graphs as causal models 42
5.1.2. Intervention in the causal system 43
5.1.3. Intervention's effect on the outcome—the causal effect . . . 44
5.2. Identifiability of the causal effect 45
5.2.1. D-separation 46
5.3. Assuming linear causal effects 47

5.3.1. Regression for estimating linear causal effects 48
5.4. Method of instrumental variables for linear causal effects not directly identifiable 49
5.4.1. The IV estimator is consistent and asymptotically normal . 51
5.4.2. Generalization of IV to multiple instruments 52
5.4.3. Generalization of IV to multiple exposures 53

6. Mendelian randomization **55**
6.1. Mendelian randomization estimator 55
6.1.1. Finite sample bias of the Mendelian randomization estimator 56
6.1.2. Statistical power of Mendelian randomization 57
6.2. Two-sample Mendelian randomization 58
6.2.1. A simple fine-mapping strategy for a single instrument MR 60
6.2.2. Allowing for multiple instruments 61
6.2.3. TWAS-like polygenic score instruments for causal inference 62
6.3. Pleiotropy in Mendelian randomization 62
6.3.1. Determining pleiotropy in multi-instrument setting 63
6.3.2. Additional sensitivity analyses 65
6.4. Multivariable Mendelian randomization 66
6.4.1. Dealing with remaining heterogeneity in effect estimates . . 68

7. Identifying causal genes in practice **69**
7.1. Causal inference using small sample individual-level data (**Ref. I**) 69
7.1.1. Novel likelihood-based model selection approach to prioritize putative causal genes 69
7.1.2. The importance of triangulation of causal evidence 71
7.1.3. Other contributions to the field 71
7.2. Genes in 16p11.2 BP4-BP5 CNV region with a causal effect on age at menarche (**Ref. II**) 72
7.2.1. Puberty timing tracks with 16p11.2 BP4-BP5 dosage 72
7.2.2. External investigation into causal genes 74
7.2.3. Summary of our contributions to the field 75
7.3. Mendelian randomization over the transcriptome (**Refs. III, IV, V**) 75
7.3.1. Practical considerations of transcriptome wide analysis . . 76
7.3.2. Improving upon existing approaches to implicate novel causal gene-trait relationships (**Ref. III**) 76
7.3.3. Sex-specific effects (**Ref. IV**) 77
7.3.4. Reverse causation: from traits to expression (**Ref. I, V**) . . 78
7.3.5. Our contributions to the field 80

8. Conclusion **81**
8.1. In terms of teaching potential 81
8.2. In terms of scientific research 82
8.2.1. Future directions 83

Appendix A. Derivation of the multivariable Mendelian randomization standard error **85**

Bibliography **89**

Acknowledgement **106**

Sisukokkuvõte (Summary in Estonian) **108**

Publications **111**

C-reactive protein upregulates the whole blood expression of *CD59* – an integrative analysis 113

Leveraging biobank-scale rare and common variant analyses to identify *ASPHD1* as the main driver of reproductive traits in the 16p11.2 locus 135

Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits 163

The role of gene expression on human sexual dimorphism: too early to call 177

Causal Inference Methods to Integrate Omics and Complex Traits. . . . 195

Curriculum Vitae **213**

Elulookirjeldus (Curriculum Vitae in Estonian) **216**

LIST OF FIGURES

1. Central dogma of molecular biology. 22
2. The formation of offspring DNA is a random process involving meiosis in both mother and father. 24
3. Manhattan plots depicting strengths of association of genetic variants with age at menarche (top) and INO80E expression (middle) in the 16p11.2 region (bottom). 34
4. Example causal graphs with and without interventions. 44
5. Method of instrumental variables for estimating the causal effect in the presence of unobserved confounders of the exposure-outcome relationship. 50
6. Statistical power analysis of the Mendelian randomization estimator. 59
7. Possible explanations for observing a significant effect in a Mendelian randomization analysis with a single instrument G. 63
8. Sensitivity analyses to detect horizontal pleiotropy in a Mendelian randomization study. 65
9. Schematic of multivariable Mendelian randomization with an outcome Y, k exposures $(X_1, X_2, \ldots, X_k)$, and m genetic instruments $(G_1, G_2, \ldots, G_m)$. 67
10. Loss of function deletions and copy gain duplications in gnomAD data covering at least 1000 base pairs in the 16p11.2 BP4-BP5 CNV region from $29.6 - 30.2$ Mb (bottom), and genes in the same region (top). 73
11. Scatter plot of $\hat{Z}$-statistics from the EGCUT-based correlation analysis (**Ref. I**) on x-axis and $\hat{Z}$-statistics from TWMR analysis (**Ref. III**) on y-axis. 79

LIST OF ABBREVIATIONS

2SLS	two stage least squares
BP	breakpoint
bp	base pair
CLPP	colocalization posterior probability
CLT	central limit theorem
CNV	copy number variant
CRP	C-reactive protein
CVD	cardiovascular disease
DNA	deoxyribonucleic acid
dpf	days post fertilization
EGCUT	Estonian Genome Centre at the University of Tartu
EGFP	enhanced green fluorescent protein
eQTL	expression quantitative trait loci
GnRH	gonadotropin-releasing hormone
GTEx	Genotype-Tissue Expression project
GWAS	genome-wide association study
IV	instrumental variable
IVW	inverse-variance weighted
kb	kilobase
LD	linkage disequilibrium
LIE	law of iterated expectations
LLN	law of large numbers
log-OR	logarithm of odds ratio
Mb	megabase
ML	maximum likelihood
MR	Mendelian randomization
mRNA	messenger RNA
MSE	mean squared error
MVMR	multivariable Mendelian randomization
OLS	ordinary least squares
PC	principal component
RCT	randomized controlled trial
RNA	ribonucleic acid
RNA-seq	RNA sequencing
SMR	summary data-based Mendelian randomization
SNP	single nucleotide polymorphism
T1E	type 1 error
T2D	type 2 diabetes
Tg	transgenic
TSS	transcription start site
TWAS	transcriptome-wide association study
UKBB	UK Biobank

LIST OF ORIGINAL PUBLICATIONS

Publications included in the thesis

This dissertation is based on the following original publications which are referred to in the text by Roman numerals (**Ref. I** to **Ref. V**):

I **Kaido Lepik**, Tarmo Annilo, Viktorija Kukuškina, eQTLGen Consortium, Kai Kisand, Zoltán Kutalik, Pärt Peterson, and Hedi Peterson. C-reactive protein upregulates the whole blood expression of CD59 - an integrative analysis. *PLoS Comput. Biol.*, 13(9):e1005766, September 2017. [1]

My contribution: Participated in conceiving and designing the study as well as framing research questions, chose the methodology, preprocessed gene expression data, performed the analyses, prepared the figures and tables, interpreted the results, and wrote the manuscript.

II Katrin Männik, Thomas Arbogast, Maarja Lepamets, **Kaido Lepik**, Anna Pellaz, Herta Ademi, Zachary A Kupchinsky, Jacob Ellegood, Catia Attanasio, Andrea Messina, Samuel Rotman, Sandra Martin-Brevet, Estelle Dubruc, Jacqueline Chrast, Jason P Lerch, Lily R Qiu, Triin Laisk, The 16p11.2 European Consortium, The Simons VIP Consortium, The eQTLGen Consortium, R Mark Henkelman, Sébastien Jacquemont, Yann Herault, Cecilia M Lindgren, Hedi Peterson, Jean Christophe Stehle, Nicholas Katsanis, Zoltan Kutalik, Serge Nef, Bogdan Draganski, Erica E Davis, Reedik Mägi, and Alexandre Reymond. Leveraging biobank-scale rare and common variant analyses to identify ASPHD1 as the main driver of reproductive traits in the 16p11.2 locus. *bioRxiv*[1], July 2019. [2]

My contribution: I was responsible for everything related to computational methods for causal inference in this paper by participating in conceiving the analysis, choosing the appropriate methodology, applying for relevant data from the eQTLGen Consortium, performing the analysis, preparing figures and tables, interpreting and communicating the results, and writing relevant parts of the manuscript.

III Eleonora Porcu, Sina Rüeger, **Kaido Lepik**, eQTLGen Consortium, BIOS Consortium, Federico A Santoni, Alexandre Reymond, and Zoltán Kutalik. Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat. Commun.*, 10(1):3300, July 2019. [3]

My contribution: Participated in method development by deriving the standard error of the multivariable Mendelian randomization causal effect estimator, implemented its calculation, and revised the manuscript.

[1] Under consideration in *Nat. Commun.*

IV Eleonora Porcu, Annique Claringbould, **Kaido Lepik**, BIOS Consortium, Tom G. Richardson, Federico A. Santoni, Lude Franke, Alexandre Reymond, Zoltán Kutalik. The role of gene expression on human sexual dimorphism: too early to call. *bioRxiv*, April 2020. [4]

My contribution: Conducted sex-specific gene expression analyses on the Estonian Biobank data and revised the manuscript.

V Eleonora Porcu, Jennifer Sjaarda, **Kaido Lepik**, Cristian Carmeli, Liza Darrous, Jonathan Sulc, Ninon Mounier, Zoltán Kutalik. Causal Inference Methods to Integrate Omics and Complex Traits. *Cold Spring Harb. Perspect. Med.*, August 2020[2]. [5]

My contribution: Conducted the analysis that compared the overlap of correlational and causal results, prepared the relevant figure, and revised the manuscript.

The publications listed above have been reprinted in the end of the dissertation with permission of the copyright owners.

Publications not included in the thesis

VI Glen James, Sulev Reisberg, **Kaido Lepik**, Nicholas Galwey, Paul Avillach, Liis Kolberg, Reedik Mägi, Tõnu Esko, Myriam Alexander, Dawn Waterworth, A Katrina Loomis, Jaak Vilo. An exploratory phenome wide association study linking asthma and liver disease genetic variants to electronic health records from the Estonian Biobank. *PLoS One*, 14(4):e0215026, April 2019. [6]

My contribution: Adapted statistical methodology, performed all the analyses, prepared the figures and tables, communicated the results, and wrote relevant parts of the manuscript.

VII Maarja Lepamets, **Kaido Lepik**, Tuuli Jürgenson, Mart Kals, Cristian Carmeli, Annique Claringbould, Murielle Bochud, Silvia Stringhini, Cisca Wijmenga, Lude Franke, Reedik Mägi, Zoltan Kutalik. New CNV quality score enables discovering novel phenotype associations from genome-wide CNV analysis. *In preparation*. [7]

My contribution: Prepared gene expression and methylation data, participated in performing the analyses, conceived and derived the quality score simulation framework, supervised the simulation analysis, interpreted the results, and wrote relevant parts of the manuscript.

[2]Early access release date

1. INTRODUCTION

A goal without a plan is just a wish

Antoine de Saint-Exupéry

One of the major human achievements of the past century has been the dramatic global rise in average life expectancy which has surged from ~50 to ~80 years in most developed countries [8,9]. Before 1950s, the increase in life expectancy was mainly fuelled by decreasing death rates at younger ages while recent advancements have come from decreasing mortality of the elderly (65+ years) [10, 11]. Old age-specific health as measured by disease-free lifespan has improved together with life expectancy but not equally as fast [8]. A growing concern is the rising burden at later stages of life due to chronic inflammatory diseases such as cardiovascular diseases (CVD) or type 2 diabetes (T2D) [11]. While environmental and lifestyle factors have an important role in the prevention or development of common diseases, research from twin studies suggests there is also a major heritable component with genetics estimated to contribute roughly half of the variability of complex traits [12]. Understanding how genetic variation translates to disease is likely to create unique possibilities for developing efficient clinical interventions to tackle or at least relieve the disease burden at older ages, and improve the quality of human life. Converting genetic information to medical applications is thus an attractive prospect in modern medicine.

1.1. Applying genotyping in medical genetics

Sequencing of the first human genome in 2001 [13, 14] and subsequent advancements in the sequencing and genotyping technologies laid a foundation for a revolution in the field of human genetics. International initiatives in the form of the HapMap Project [15] and the 1000 Genomes Project [16] were taken on by research communities worldwide to systematically catalogue and provide a reference for human genetic variation. These resources facilitated genome-wide association studies (GWAS) which emerged as powerful tools for uncovering the genetic basis of complex traits [17]. More than 24,000 associations between phenotypic traits and genetic variation across the genome from more than 2,500 publications were reported from March 2005 to September 2016 in the GWAS Catalog, an expertly curated database of published GWAS associations satisfying quality control criteria [18]. While intervention on disease-associated genetic regions (loci) at a single nucleotide resolution could be theoretically possible with genome editing techniques, it is currently infeasible for several reasons. First, the identification of causal loci has proven difficult as close-by genetic regions tend to share association signals due to genome inheritance patterns [19]. Second, most common diseases are highly polygenic [20] (possibly omnigenic [21]) and it would be difficult to modify thousands of genomic loci.

At the moment, genotyping in clinical setting is mostly useful for diagnosing monogenic diseases and facilitating cure through gene therapy. It has also been successfully applied in pharmacogenetics where it is often the case that metabolization of certain drugs is controlled by only a few genes (e.g. Very Important Pharmacogenes [22]); based on their genotype, patients can be subscribed drugs that are likely to work well on them and not cause adverse reactions [23]. The situation is more difficult with polygenic traits, though GWAS have contributed to genetics-informed personalized and predictive medicine approaches. Especially popular are polygenic risk scores (PRS) [24] which aggregate the contribution of many genetic loci toward disease progression and enable disease risk prediction at an early age. While promising in terms of clinical useability, PRSs are currently plagued by population-specific biases and inconsistencies of individual risk estimates [25–28]. Either way, designing novel intervention strategies for most common diseases is challenging due to polygenicity. Interventions would invariably revolve around elimination of disease risk factors such as reducing stress, having a healthy diet, exercising regularly, not smoking, etc.—all of which should be pursued by knowledgeable individuals regardless of their genetic susceptibility to any disease. This limits the usefulness of genotype-based risk prediction to informed regular screening of high-risk individuals.

1.2. Identifying intervention candidates for complex diseases

In effect, disease intervention strategies could rather concentrate on downstream functional consequences of genetic variation. At the same time, it is not enough to identify genes or gene products that are merely correlated with diseases or quantitative traits of interest. While these could sometimes point in the right direction, associations could also emerge purely due to confounding factors which independently affect both the intermediate trait and the final outcome [29–31]. Thus, the focus should be on finding intermediate traits that causally affect the disease: gene expression or other molecular traits even further downstream that participate in biological processes leading to disease development. Doing so could facilitate intervention by developing drugs that target said traits [32–34].

The preferred standard for causal inference is the randomized controlled trial (RCT), often utilized in clinical research or pharmacology to test the efficacy of medical interventions such as new drugs [35]. RCTs work by allocating individuals into two groups based on intervention strategy—a treatment group that is subjected to intervention, and a control group that is not—which are subsequently compared in terms of clinical outcomes. Randomization into groups eliminates selection bias and confounding, thus enabling causal reasoning [35]. Unfortunately, undertaking RCTs is expensive, time-consuming and in some cases unethical, leading to a narrowness of scope [36]. There is a need for generating candidate intervention strategies fast and in bulk. This is where bioinformatics and computational biology enter the fray [37].

In an effort to understand all the intermediate steps from genetic variation to final disease outcome, working with multiple big biological datasets—genome (genetic variations), transcriptome (gene expression levels), proteome (protein abundances), phenome (phenotypic trait values) and other data—is necessary. Such *omics* datasets have been increasing at a tremendous pace due to great work by large national population-based biobank initiatives, including the Estonian Biobank of the Estonian Genome Centre at the University of Tartu (EGCUT) [38]. To extract novel information out of the data, integration of these datasets and the use, modification or development of new state-of-the-art statistical methods are needed [39]. Integrative analysis of omics data constitutes a broad field of study; in this dissertation, I will concentrate on its subfield—methods facilitating causal inference. I will introduce both existing approaches and our own developments. Due to the advanced state of transcriptome studies compared to other intermediate omics layers between genomics and phenomics, most of the research in this area has focussed on identifying trait-associated causal genes. This will also be our focus, though the ideas we cover can be applied to other omics data as well.

1.2.1. Identifying trait-associated causal genes

Teasing out causality based on statistics and computational methods alone is difficult and invariably rests on several assumptions, some of which are impossible to verify in practice [40]. Luckily, genotype data provides an anchor for teasing out causality. As genetic variants are randomly inherited and fixed at birth, there can be almost no confounding factors to genetic associations (see Subsection 2.2.1 for some exceptions), and no reverse causation (SNP-trait associations could not be caused by the trait) [30].

About 90% of disease-associated genetic variants lie outside the coding region of genes [41]. These variants have no control on the type of protein encoded by the gene, rather they influence the amount of protein it produces through regulation of its expression levels [42]. Unfortunately, mapping GWAS-significant variants to functional genes is complicated. It is difficult even to determine causal variants due to shared association signals, though this can be attempted with fine-mapping methods (in-depth analysis focused on determining causality in specific well-defined genomic regions) [43, 44]. A naive approach would link these variants to causative genes based on closest proximity but it is not always so straightforward [45, 46]. Instead, we could hypothesize that understanding the biological processes leading to complex traits can be aided by overlapping the genetic basis of gene expression and complex trait variability. As complex trait-associated genetic variants also tend to be involved in gene expression regulation [41, 42, 47], we can ask whether the same loci underlies both the intermediate and final traits. While overlap in itself does not mean causality (there are so many associated variants that overlaps could happen even by chance), there are computational approaches for gene prioritization based on this premise, termed colocalization

analyses [48, 49]. Another group of methods designed to identify trait-associated genes is called transcription-wide association studies (TWAS) [50–52]. These methods use genotype data to estimate the level of gene expression that is attributable to genetics and then test for its association with complex traits. If followed by gene-level fine-mapping, TWAS genes can be prioritized further [53].

Colocalization and TWAS approaches are blind to the strength and direction of potential causal effects, vulnerable to confounding, and unable to discriminate causality from pleiotropy [5]. A group of methods designed to overcome these limitations and formally test for a causal effect between a modifiable exposure (gene expression) and an outcome (complex disease) is called instrumental variable (IV) analysis, where IV is a variable with a demonstrable causal effect on the exposure. On the assumption that the IV can influence the outcome only through the exposure, the exposure has a causal effect on the outcome if and only if the IV is associated with the outcome. In epidemiology and statistical genetics, IVs are genetic variants and the IV analysis method is named Mendelian randomization (MR) after Mendel's laws of inheritance [30]. Effectively, MR mimicks the design of RCTs and is thus a natural computational extension to the preferred standard of causal inference. It has been successfully used to elucidate functional relationships between complex traits and diseases, such as the causal role of lipids—particularly the detrimental effect of low-density lipoprotein cholesterol—on CVD risk [54–56].

1.3. Aims of the dissertation

There are two major goals for this dissertation. First, to provide an overview of the field of causal inference in statistical genetics with the focus on causal gene discovery. I will attempt to bring together and provide a uniform treatment of related concepts from traditional statistics, econometrics, genetics, and formal theory of causality. Popular methods for statistical gene prioritization—both association- and causality-based—will be introduced and placed in context. While the underlying theory of the concepts and methods covered (in chapters 2-6) is known, it is scattered between several fields of science where related matters are not necessarily approached in the same way. I believe there is great value in integrating similar research under the same framework, specifically to further understanding. To the best of my knowledge, this has not been done before in statistics-based causal gene discovery. The treatment and interpretation of the material presented in this dissertation is thus entirely my own. I have done my best to approach the topics in a structured way and from the ground up, ideally to provide a useful teaching material and resource for the scientists interested in statistical genetics, perhaps given as a graduate level course in the future.

Second, I will tackle functional genomics in applied settings by tweaking and utilizing causal inference methods to prioritize candidate genes with a demonstrable causal effect on complex traits and diseases—with the purpose to help bridge

the gap between pharmaceutical drug discovery and computational approaches for target identification. I will cover my research and results from **Refs. I-V** (in Chapter 7) which have been published in scientific journals and reprinted at the end of the dissertation. All of the theory covered previously in the dissertation will serve to make these research contributions understandable and relatable. It further enables me to be succinct, to the point and purposefully concise in covering my research results (the full papers are published and easily accessible at the end of the dissertation).

In particular, the aims of the dissertation are the following:

1. Provide a uniform treatment of gene prioritization methods in the confines of causality.
2. Develop and exploit causal inference methods to interrogate causal relationships between gene expression and complex traits (**Refs. I-V**).
   (a) Elucidate the cause and effect of gene expression on C-reactive protein (CRP) in order to explore its functional role in healthy aging and inflammatory processes (**Ref. I**).
   (b) Identify causal genes mediating sexual development in the 16p11.2 breakpoint (BP) 4 to 5 copy number variable (CNV) region (**Ref. II**).
   (c) Adapt causal inference methodology to detect causal genes for complex traits over the entire transcriptome (**Ref. III**).
   (d) Investigate whether sex-specificity in complex traits can be attributed to sex-specificity in gene expression regulation, with the focus on uncovering sex-specific causal genes (**Ref. IV**).
   (e) Show that trait-associated genes tend not to entail causal genes and instead could overlap more with trait-affected genes (**Ref. V**).

I start off by introducing the fundamental concepts of biological functioning of living organisms in Chapter 2 and statistical genetics in Chapter 3. In Chapter 4, I will cover the popular gene prioritization methods discussed above—fine-mapping, colocalization and TWAS approaches—in more rigorous detail. In Chapter 5, I will formalize the concept of causality and explore both the challenges and opportunities within. In Chapter 6, I will introduce different flavours of MR together with our multivariable MR approach for identifying trait-associated causal genes (**Refs. II, III**). Finally, I will discuss our research results in Chapter 7 (**Refs. I-V**) and underline the importance of triangulation of evidence for making causal claims.

The dissertation is structured in a way that allows to harmonize related concepts of statistical genetics. Hence, later chapters have abundant references to the results derived previously.

2. FUNDAMENTALS OF BIOLOGICAL PRINCIPLES

> *Speak English!* said the Eaglet. *I don't know the meaning of half those long words, and, what's more, I don't believe you do either!*
>
> Lewis Carroll *(Alice in Wonderland)*

Identifying causal mechanisms in biological systems necessitates basic knowledge about the principles of life and the machinery that keeps it going. After all, acquiring understanding of the human organism and its functions inevitably needs to precede any attempt to manipulate disease processes. Only with sufficient information on the mechanics of disease can pharmaceuticals hope to develop intervention strategies. Domain knowledge is equally important for statistical analyses that are to help in determining causal relationships. In order to attain the required level of domain proficiency, I will introduce in this chapter the fundamental concepts of biological functioning of life.

2.1. Cells maintain life using proteins

Cells are the smallest units capable of independent function in any living organism [57]. A human organism represents a complex coordinated system of trillions of cells, all responsible for specific tasks to keep it operational [58]. Cells can be grouped into different types depending on their function (e.g. neutrophils, basophils, eosonophils and lymphocytes are all types of white blood cells responsible for various aspects of immune response [1]). Groups of similar cells—possibly from different but closely related types—that carry out specific functions are called tissues (e.g. whole blood consisting of red blood cells, white blood cells and platelets is responsible for transporting vital substances between other tissues while hypothalamus—a collection of tissues in the brain—coordinates sexual development [2]). Each tissue is thus a mixture of different cell types with varying characteristics. To understand the working mechanisms of cells and tissues, we need to delve into genetics.

2.1.1. Governed by the central dogma of molecular biology

The function of each cell is encoded in genetic material called the genome. The latter consists of deoxyribonucleic acid (DNA) molecules (chromosomes) composed of two complementary strands of nucleotides joined together into a double helix. Each nucleotide in a DNA molecule is composed of one of four different nucleobases: A – adenine, T – thymine, G – guanine, C – cytosine. These are the primary building blocks of DNA. The complementary nature of DNA strands is guaranteed by hydrogen bonds forming only between adenine and thymine, or guanine and cytosine. As a consequence, both strands taken separately contain the same biological information—having one reveals the other. Human cells are

diploid which means that each chromosome comes with a copy, one from each parent. There are a total of 23 pairs of chromosomes in every human cell: 22 pairs of autosomes and 1 pair of sex chromosomes. An individual set of chromosomes contains approximately 3 billion AT and GC base pairs (bp) stuck onto each other in sequences [13, 14]. Decoding for the information hidden in these sequences holds the key to understanding the functioning of life and represents arguably the greatest challenge in human research.

Each cell in the human organism, irrespective of type, contains the same genetic makeup. The interpretation of this information comes down to the work of genes—sequences of DNA that encode functional products such as proteins. Different patterns of gene expression—the process of synthesizing gene products—is the reason why cell types present physiological irregularities [57]. Gene expression starts with transcription of DNA into ribonucleic acid (RNA). An enzyme RNA polymerase binds to a specific DNA sequence immediately upstream of the gene called promoter to initiate transcription. Hydrogen bonds between complementary nucleobases in the DNA double helix are broken down and one of the DNA strands is used as a template to create a complementary RNA strand instead: a G, C and A is attached to every C, G and T in the template, respectively, but thymine is replaced by a less stable nucleobase uracil (U) to pair with A.

If the expressed gene encoded a protein, the newly created RNA molecule is called messenger RNA (mRNA) and needs to be synthesized into the actual protein product in a translation event. This process is orchestrated by a macromolecule called ribosome that binds to the mRNA strand, locates the starting sequence and proceeds to attach amino acids onto groups of three adjacent nucleotides (codons) in the strand. The ribosome continues to match nucleotide triplets with specific amino acids until it reaches a stopping sequence. The chain of amino acids that is created as a result of translation is exactly what makes up a protein, or peptide if the chain is short.

In essence, cells need to execute functions to sustain life. These functions are carried out by gene products. The commands to produce gene products are encoded in the DNA which crudely serves as a cookbook. Transcription events create copies of the commands (genes/recipes) which translation events convert to proteins. Once the genetic information has transferred from DNA through RNA to making a protein, it cannot be reversed. This principle is captured in the central dogma of molecular biology (Figure 1).

2.1.2. Proteins' involvement in disease through gene expression

Proteins are the workhorses of life by having any of myriad of responsibilities from keeping the structure of cells to transmitting signals between them [57]. To grasp the magnitude of proteins' involvement in different functions of the organism, consider that the human genome has approximately 20,000 protein coding genes [59]. The number of unique proteins that could circulate in a human or-

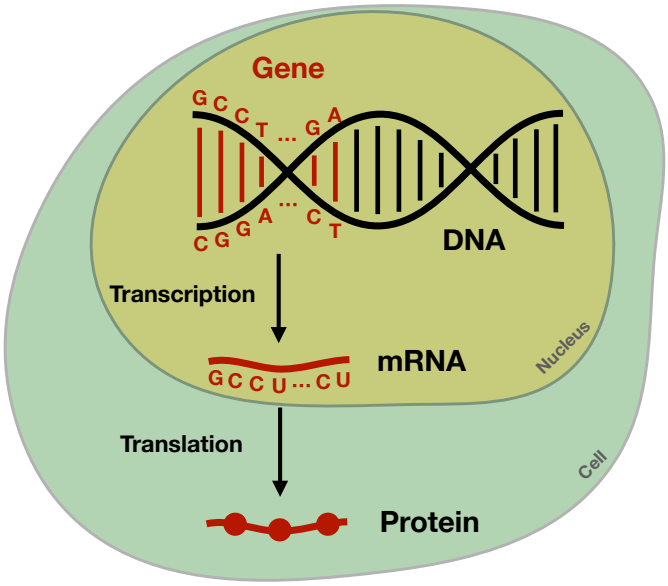


Figure 1: Central dogma of molecular biology. Specific DNA sequences called protein coding genes are copied into mRNAs which in turn are synthesized into proteins. This flow of genetic information cannot be reversed. Proteins are responsible for a vast array of functions in the organism, thus the central dogma of molecular biology can be characterised by genetic variation leading to downstream biological variability.

ganism is even higher because genes can actually be transcribed to produce different RNA sequences—resulting in different proteins—through what is called alternative splicing. Even this process is regulated by proteins. It is therefore not surprising that some protein activity can also drive disease progression. Modern drug-based medicine is built on this premise. Indeed, the majority of approved drugs on the market today target proteins by manipulating their function in some way or another deemed counteractive to disease [60].

The first step to drug-based treatment of diseases is identifying proteins that are involved in disease processes. Pharmaceuticals are constantly in the hunt for new targets but this can be a slow and expensive undertaking [34]. Bioinformatics approaches seek to relax these barriers through reliance on computation and statistics instead of manual lab work. Observational data on protein levels is needed for that purpose. Unfortunately, this data is lacking due to technological limitations in ascertaining protein abundances [61, 62]. Fortunately, the central dogma of molecular biology ensures that human biological variability depends on upstream transcriptomic variability. Furthermore, gene expression measurements are much more attainable as the similarities between RNA and DNA (Subsection 2.1.1) enable to use the same technology developed for DNA sequencing approaches to sequence RNA as well [62]. Even though variations inside the coding region of

genes could affect protein structure while not affecting circulating mRNA levels at the same time, gene expression represents still a good proxy to protein levels as evidenced by the central dogma (Figure 1) [61]. In this dissertation, for these reasons, we are looking for genes with expression levels causally related to complex traits to gain insight into disease processes. To facilitate statistical analysis in this direction—invariably based on observational data from samples of individuals—we need to introduce the basic vocabulary of genetics with an emphasis on heredity in living organisms.

2.2. The basics of genetics

More than 99.9% of the genome of any two individuals is estimated to be exactly identical [13]—after all, everyone looks and functions largely the same. The positions in the genome with bp differences are said to exhibit genetic variation. If only a single bp is involved, the position is referred to as a single nucleotide variation (SNV), or single nucleotide polymorphism (SNP) if the substitution is frequent in a population (at least 1%). Single bp differences of DNA make up around 90% of all human genetic variations with tens of millions frequent enough to be considered SNPs [63]. Mutations in the genome can result in bp insertions or deletions which can cover larger stretches of DNA sequence at a time. If the insertion is a repeat of an existing sequence, it is called a duplication. Variations in the number of repeats of a DNA sequence, such as due to duplication or deletion events, are referred to as copy number variations (CNV). The number of bp these cover could reach into thousands (measured in kilo base pairs (kb), each 1 000 bp) or millions (measured in mega base pairs (Mb), each 1 000 kb), potentially resulting in serious consequences to the functioning of the organism.

2.2.1. DNA inheritance patterns

An organism's DNA is made up of genetic sequences passed down by their parents. DNA inheritance can be characterized by the formation of reproductive cells (gametes) in the process called meiosis, and the fusion of maternal and paternal gametes during fertilization (Figure 2). During meiosis in diploid germ cells of parents, a copy is made of each chromosome which is attached to its original counterpart, forming a pair of chromatids. The chromatids of homologous chromosomes can exchange genetic material in a process called crossing over before dividing into separate gametes, making four haploid cells in total. Finally, two gametes together with their genetic material, one from each parent, are fused into a diploid cell from which the new individual develops. The offspring's DNA is thus a random combination of the parents' DNA where the randomness originates from both chromosomal crossover and selection of gametes to be fused. The non-deterministic nature of DNA inheritance introduces variability into human phenotypes, facilitating evolution. It is also the basis for attempting causal inference with statistical methods (Chapter 6).

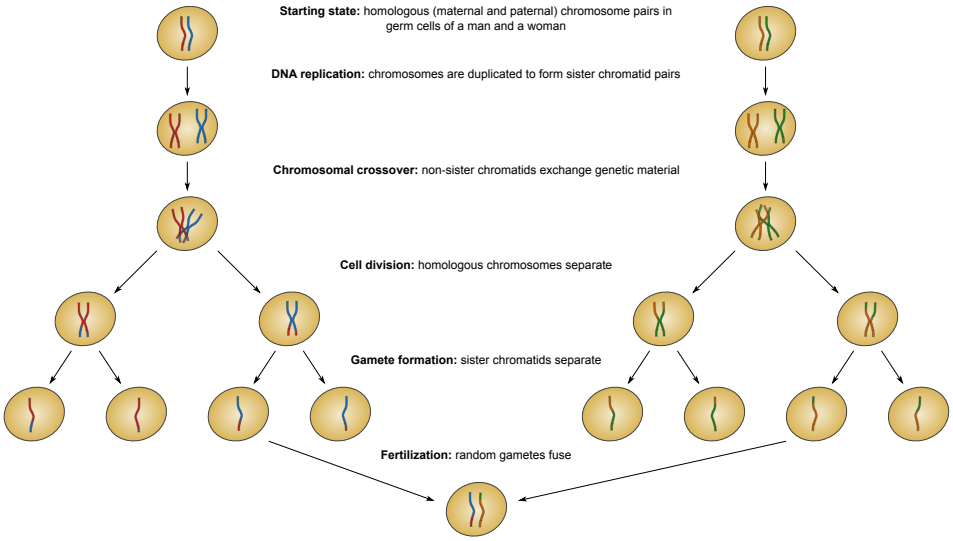


Figure 2: The formation of offspring DNA involving meiosis in both mother and father to produce gametes which are subsequently fused during fertilization. This is a non-deterministic process where the randomness is inherent to crossing over (recombination of genetic information between non-sister chromatids of homologous chromosomes) and the determination of gametes to be fused together.

The biological mechanism of reproduction is a closed system but outside influences could still affect DNA inheritance patterns. For example, people are known to be more likely to mate with other people within the same geographical region (giving rise to population stratification [64, 65]) who furthermore have similar phenotypic characteristics to themselves (a phenomenon called assortative mating [66]). Population-based samples are not random for these reasons, exhibiting regularities in geno- and phenotype values (differences between and similarities within subpopulations) which statistical methods need to account for. Furthermore, there are likely no more than a few crossover events per human chromosome (about 1.6 on average) [67]. As a result, parents' genomes are passed on to offspring in large chunks with close-by genetic loci likely to be inherited together. This introduces linkage disequilibrium (LD)—dependence between alleles at different loci in a population—which makes statistical analysis-based interpretation of genetic underpinnings of complex traits particularly challenging.

2.2.2. Genetic architecture of complex traits

Many different genes can influence the same phenotypic characterizations (traits) of the human organism (e.g. disease states) [20]. Such polygenic traits cannot be described by variation in a single genomic region and are deemed complex as a result. By the omnigenic theory, some traits may even be influenced by the majority if not all genes in the genome [21]. The reverse also holds—any one gene can influence multiple seemingly unrelated traits, a phenomenon called

horizontal pleiotropy. The mechanisms of pleiotropy incorporate instances of a gene product having various functions (e.g. in separate tissues) or regulating the expression of several genes but also include alternative splicing [68]. Pleiotropy, like polygenicity, is ubiquitous among human complex traits [17].

Genes act on traits by synthesizing functional products, a process governed by the underlying DNA (Subsection 2.1.1). This notion has prompted many studies to investigate the effect of genetic variation on trait development in genome-wide analyses. It turns out that the majority (about 90%) of genomic loci associated to complex traits do not fall into the coding region of genes and thus do not affect protein structure [41]. Instead, these loci are enriched in regions responsible for regulation of gene expression levels and thus can affect the amount of protein produced [42].

A genomic locus that is associated to gene expression levels is called an expression quantitative trait locus (eQTL). It is referred to as *cis*-eQTL if it has a direct effect on the gene, otherwise *trans*-eQTL. As such, *cis*-eQTLs have larger effects on gene expression and tend to be located close to transcription start sites (TSS) of genes (e.g. within 1 Mb). A good example is a SNP influencing the activity of a transcription-regulating protein (transcription factor) that either promotes or suppresses the binding of RNA polymerase to mRNA. *Trans*-eQTLs tend to be more distant from respective genes, locating even on different chromosomes. Following the previous example, a SNP that controls the production of a transcription factor in *cis* is *trans*-acting on the gene that this protein affects.

Due to polygenicity, genes across the genome tend to have only tiny individual effects on the development of complex traits. This makes intervention on disease complicated. However, there could be a small number of core genes with larger and more direct effects that other genes are acting upon [21]. This would be indicative of elaborate gene regulatory networks and pathways underlying complex traits. Indeed, unlinked trait-associated genetic variations often have *trans*-effects on the same genes known to play important roles in disease aetiology [69]. Modulating the activity of a few central genes holds a lot more promise in terms of disease intervention strategies. Methods of statistical genetics are used to tease these genes out.

3. FUNDAMENTALS OF STATISTICAL GENETICS

> Give me six hours to chop down a tree and I will spend the first four sharpening the axe
>
> Abraham Lincoln

In the modern day and age, statistical analyses are becoming ever important for scientific discoveries in many disciplines, and epidemiology is no exception. With tens of thousands of genes and countless number of ways these could interact with each other or with the environment, teasing out disease-relevant genes with functional characteristics invariably benefits from the greatly increased speed and simplicity that computational methods provide for testing hypotheses in bulk. Not only is the rate of acquiring new knowledge increased with computational analyses, exploiting what we have learned to unravel yet more information works as an exponential. However, reaching to correct conclusions in all this necessitates sufficient understanding of the inner-workings of the tools and methods used to produce the results. To facilitate this understanding, we will first lay the groundwork in terms of basic and fundamental concepts in statistical genetics; this will pave the way for grasping the ideas behind more complex methodology introduced later in the dissertation.

3.1. Genome wide association study

Let Y and G_i, $i \in \{1,2,\ldots,m\} =: \mathcal{I}_m$ be random variables of a complex trait and genetic variants (SNPs), respectively. A GWAS is about estimating the effect size β_i of G_i on Y, most often pursued using generalized linear models

$$g(\mu) = \beta_0 + \sum_{i=1}^{m} G_i \beta_i,$$

where $\mu = \mathrm{E}(Y \mid G_1, G_2, \ldots, G_m)$ and $g(\cdot)$ is a link function. An identity link $g(\mu) = \mu$ is often used for a quantitative Y and a logit link $g(\mu) = \ln\left(\mu(1-\mu)^{-1}\right)$ for a binary Y. If Y is a gene expression trait, then the association study is called an eQTL analysis.

In a finite sample of size n, observations of two SNPs G_i and G_j ($i \neq j$) could be exactly identical due to LD, resulting in ill-defined effect estimates. Another problem is that usually $n \ll m$ and effect estimates are not uniquely defined. A workaround is to test every SNP G_i independently of other SNPs. As covariates such as age, sex and genotype principal components (PCs) are often used to decrease the variability of Y and reduce residual confounding (due to population stratification), the following model is used in practice:

$$g(\mathrm{E}(Y \mid G_i, \boldsymbol{U})) = \beta_0 + G_i \beta_i + \boldsymbol{U}' \beta_U, \tag{3.1}$$

where $\boldsymbol{U} = (U_1, U_2, \ldots, U_p)'$ is a random vector of p covariates and β_U is a vector of covariate effects. Though generalized linear models are popular due to computational speed, low memory requirements and ease of use, genotype PCs are only proxies for population stratification and do not help against cryptic relatedness [70, 71]. Generalized linear mixed models

$$g(\mathrm{E}(Y \mid G_i, \boldsymbol{U}, \boldsymbol{G}_{-i})) = \beta_0 + G_i\beta_i + \boldsymbol{U}'\beta_U + \boldsymbol{G}'_{-i}w, \tag{3.2}$$

where $\boldsymbol{G}_{-i}$ is a random vector of genetic variants without G_i and $w \sim \mathcal{N}(0, \sigma_w^2 I)$ is a vector of corresponding random effects, provide better control for these confounding factors and recent methodological advancements have also brought them into more widespread use [72, 73].

Parameters in GWAS models are estimated from a random sample through an optimization procedure. For every SNP G_i, the outcome of the operation is an effect size estimate $\hat{\beta}_i$ and its standard error $\hat{\sigma}_i$, allowing to test for a null hypothesis $H_0 : \beta_i = 0$ with a Wald test. This is a general theory and applies to both of the models specified in equations 3.1 and 3.2 above, irrespective of whether the outcome Y is quantitative or binary. The differences come in the form of parameter estimates. Since ordinary linear regression with a quantitative Y is the most fundamental method and also the most relevant in terms of the research in this dissertation, we will treat it in slightly more detail in the following.

3.1.1. Ordinary least squares estimator

Consider a GWAS with a quantitative complex trait Y using the linear model 3.1 above. For convenience, let us group all the independent variables together with the intercept into a single random vector $\boldsymbol{X} : k \times 1$, such that

$$Y = \beta_0 + G_i\beta_i + \boldsymbol{U}'\beta_U + \varepsilon = \boldsymbol{X}'\beta_X + \varepsilon, \tag{3.3}$$

where $\beta_X : k \times 1$ is the corresponding vector of effect sizes and ε denotes for random fluctuations in the outcome Y not captured by variables in $\boldsymbol{X}$. In terms of sample realizations within an ordinary linear regression framework, we can write this model as

$$y = X\beta_X + \varepsilon\,, \qquad \begin{matrix} \mathrm{E}(\varepsilon \mid X) = 0 \\ \mathrm{Var}(\varepsilon \mid X) = \sigma_\varepsilon^2 I \end{matrix}, \tag{3.4}$$

where $y : n \times 1$ is a vector of complex trait observations, $X : n \times k$ is a design matrix and errors $\varepsilon : n \times 1$ are assumed to be uncorrelated with conditional mean 0 and finite variance σ_ε^2. The well-known estimator of β_X which minimizes the sum of squared residuals—the ordinary least squares (OLS) estimator—is the following linear combination of y:

$$\begin{aligned} \hat{\beta}_X &= \underset{\beta_X}{\operatorname{argmin}} (y - X\beta_X)'(y - X\beta_X) \\ &= (X'X)^{-1}X'y\,. \end{aligned} \tag{3.5}$$

Considering X as fixed is the usual assumption in practice which makes it particularly easy to find the expected value and variance of $\hat{\beta}_X$:

$$\mathrm{E}\big(\hat{\beta}_X \mid X\big) = \beta_X, \tag{3.6}$$

$$\mathrm{Var}\big(\hat{\beta}_X \mid X\big) = \sigma_\varepsilon^2 (X'X)^{-1}. \tag{3.7}$$

The OLS estimator $\hat{\beta}_X$ is thus conditionally unbiased—it is easy to see through the application of law of iterated expectations (LIE) that it is also unconditionally unbiased—and according to the Gauss-Markov theorem, it is optimal in the sense that it has the minimal variance of all other linear unbiased estimators. Though the variance cannot be calculated exactly if σ_ε^2 is unknown, the latter can be estimated by the sample variance of the residuals as follows:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n}(y - X\hat{\beta}_X)'(y - X\hat{\beta}_X). \tag{3.8}$$

Note that we did not enforce any distributional assumptions on the variables in model 3.4 to derive the OLS estimator together with its expected value and variance. However, to facilitate hypothesis testing in finite samples, it is usually assumed that errors are normally distributed. In that case, $\hat{\beta}_X$ as a linear transformation of a normally distributed random variable is itself normally distributed:

$$\hat{\beta}_X \mid X \sim \mathcal{N}\left(\beta_X, \sigma_\varepsilon^2 (X'X)^{-1}\right). \tag{3.9}$$

Even when the errors are not normally distributed, the relation 3.9 holds asymptotically (Subsections 3.2.2 and 5.4.1).

Normality of the parameter estimates makes it straightforward to test for the null hypothesis $H_0 : \beta_i = 0$ of SNP $i \in \mathcal{I}_m$ using the Wald test. However, the number of SNPs m to test can be very large, creating a huge multiple testing burden. Coupled with the observation that individual SNPs tend to have only tiny effects on the outcome [21], the sample size n needs to be quite large to have any reasonable chance of rejecting the null hypothesis even if the alternative really is true (see power calculations in Subsection 6.1.2 below). As a result, GWAS effect estimates from multiple samples are often combined together to improve statistical power.

3.1.2. Meta-analysis

While GWAS are usually undertaken by biobanks with population cohort data in accordance with model 3.4, these efforts are often lead centrally by international consortia with the purpose of combining together individual effect estimates for maximal statistical power. This can be done through meta-analysis.

Let $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)'$ be the effect size estimates of SNP G in k different studies with standard errors σ_i for all $i \in \mathcal{I}_k$. Each individual estimate is assumed to fluctuate around the true effect, $\mathrm{E}\big(\hat{\beta}_i\big) = \beta_*$, the extent of which is specified by

the within-study variance σ_i^2 and potentially—if studies are not directly comparable (e.g. due to ancestry) and thus measure slightly different phenomena—also by the between-study variance σ_τ^2, such that $\mathrm{Var}(\hat{\beta}_i) = \sigma_i^2 + \sigma_\tau^2$. Effect estimates can also be correlated with each other if there is sample overlap between studies. The covariance matrix of the effect size vector, $\mathrm{Var}(\hat{\beta}) = V$, can thus be any $k \times k$ symmetric positive semi-definite matrix. Furthermore, estimates gathered from published studies of smaller sample sizes can be biased [74]. Considering all of the above, we have the following model for the relationship between the true effect β_* and estimates $\hat{\beta}$:

$$V^{-\frac{1}{2}}\hat{\beta} = \mathrm{bias} + \left(V^{-\frac{1}{2}}1_k\right)\beta_* + \varepsilon, \qquad \begin{matrix}\mathrm{E}(\varepsilon) = 0 \\ \mathrm{Var}(\varepsilon) = I\end{matrix}, \tag{3.10}$$

where $V^{\frac{1}{2}}$ is a matrix satisfying $V^{\frac{1}{2}}V^{\frac{1}{2}} = V$ (any covariance matrix V can be factorized like that through eigendecomposition) and $V^{-\frac{1}{2}}$ is its inverse, 1_k is a k-vector of ones, and bias refers to the small study (publication) bias. Note that model 3.10 belongs to the class of ordinary linear regression models 3.4, thus we can use the OLS solution 3.5 to provide the true effect β_* with an unbiased estimate with variance 3.7 which is optimal since the Gauss-Markov theorem still holds. The pertinent question remaining is how exactly to construct the covariance matrix V.

We derived model 3.10 in a very general form. Doing so helps to treat similar subjects in a coordinated manner (see also Section 6.3) but can be overly lax in some specific cases. For example, GWAS meta-analyses are often based on independent studies ($V_{ij} = \mathrm{Cov}(\hat{\beta}_i, \hat{\beta}_j) = 0$ for $i \neq j$) not exposed to subjective inclusion (bias $= 0$). Under these restrictions, the least squares optimization procedure on model 3.10 is equivalent to minimizing the sum of inverse variance (V_{ii}^{-1}) weighted (IVW) squares of residuals and the solution conveniently turns out to be the IVW average of individual effect estimates

$$\hat{\beta}_{IVW} = \frac{\sum_{i=1}^{k} V_{ii}^{-1}\hat{\beta}_i}{\sum_{i=1}^{k} V_{ii}^{-1}} \tag{3.11}$$

with accompanying variance

$$\mathrm{Var}\left(\hat{\beta}_{IVW}\right) = \frac{1}{\sum_{i=1}^{k} V_{ii}^{-1}}. \tag{3.12}$$

Note that the number of studies can be one ($k = 1$) in which case the results simply correspond to the estimates of the single cohort.

The meta-analysis can be further simplified if all studies are comparable (e.g. from the same ancestry) and thus measure the same phenomenon ($\sigma_\tau^2 = 0$). In

this case, the extent of fluctuations of each estimate around the true effect β_* is simply characterised by the within-study variance ($V_{ii} = \mathrm{Var}(\hat{\beta}_i) = \sigma_i^2$). Validity of this assumption of effect homogeneity should ideally be ascertained before the meta-analysis takes place but its plausibility can also be investigated retrospectively. Indeed, since effect estimates $\hat{\beta}_i$ were assumed to be normally distributed (relation 3.9) and independent (IVW solution 3.11), the sum of squares of k standardized deviations of the estimates $\hat{\beta}_i$ from the true effect β_* follows a chi-square distribution with k degrees of freedom. The true effect remains unknown but can be replaced by its estimate $\hat{\beta}_{IVW}$. Doing so spends one degree of freedom and leads to a test statistic called Cochran's Q [75]:

$$Q = \sum_{i=1}^{k} \frac{\left(\hat{\beta}_i - \hat{\beta}_{IVW}\right)^2}{\sigma_i^2} \sim \chi^2_{k-1}. \tag{3.13}$$

Testing for the hypothesis of no heterogeneity ($H_0 : \sigma_\tau^2 = 0$) is equivalent to testing whether Q is significantly different from its expected value $k-1$. Excess values of Q are indicative of $V_{ii} = \mathrm{Var}(\hat{\beta}_i) = \sigma_i^2 + \sigma_\tau^2$ being a more plausible decomposition of the variance of $\hat{\beta}_i$ around β_*. The between-study variance σ_τ^2 can be estimated from this excess, provided it is brought to the same scale with within-study variances (for details, see the DerSimonian and Laird method [76]). However, estimates $\hat{\beta}_i$ deviating too much from the expected normal distribution $\mathcal{N}(\beta_*, \sigma_i^2)$ could also be eliminated from the analysis as outliers.

If small study (publication) bias cannot be ruled out then the intercept in model 3.10 could be non-zero and should thus be allowed. The model with the intercept is called Egger regression [74] and provides a bias-free estimate of the true effect β_*.

3.2. Summary statistics

In GWAS, the meta-analysis is repeated for every SNP G_i, $i \in \mathcal{I}_m$. The resulting effect sizes $\hat{\beta}_i$ and variances $\widehat{\mathrm{Var}}(\hat{\beta}_i)$ are collectively referred to as GWAS summary statistics and are usually published together with corresponding Z-scores $\hat{Z}_i = \hat{\beta}_i / \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_i)}$. Associations are deemed statistically significant and subject to further examination if Z-scores are greater than some pre-specified threshold. Standard practice in GWAS is to assume 10^6 effective number of (independent) SNPs and thus require $|\hat{Z}_i| > |\Phi^{-1}(2.5 \times 10^{-8})|$, where Φ is the cumulative distribution function of the standard normal distribution.

Many downstream applications developed for gene prioritization and causal inference (among others) require only summary statistics as input. This is very useful because phenotype data is not needed in the process. Summary statistics can easily be shared—there are less privacy concerns as individuals are not easily identifiable based on summary statistics. Thus data can be reused across different research groups, for different analyses and purposes.

3.2.1. Standardization

To simplify the interpretation of effect estimates and the math involved in these methods, summary statistics—if not already calculated on standardized data—are often modified to correspond to variables with zero mean and unit variance. Let $G_i^s = s(G_i - c)$ be a random variable of a genetic variant that has been shifted and scaled by scalars c and s, respectively. Denote by β_i^s the effect of G_i^s on Y and by β_0^s the new linear regression intercept. Then model 3.3 is

$$\begin{aligned} Y &= \beta_0^s + G_i^s \beta_i^s + \boldsymbol{U}\beta_U + \varepsilon \\ &= \left(\beta_0^s - cs\beta_i^s\right) + G_i\left(s\beta_i^s\right) + \boldsymbol{U}\beta_U + \varepsilon \\ &= \beta_0 + G_i\beta_i + \boldsymbol{U}\beta_U + \varepsilon. \end{aligned}$$

As β_i corresponds to the effect size of the original genotype G_i on Y, we can express sample estimates $\hat{\beta}_i^s = \frac{1}{s}\hat{\beta}_i$ and $\hat{\sigma}_{\hat{\beta}_i^s} = \frac{1}{s}\hat{\sigma}_{\hat{\beta}_i}$. Note that $\frac{\hat{\beta}_i}{\hat{\sigma}_{\hat{\beta}_i}} = \frac{\hat{\beta}_i^s}{\hat{\sigma}_{\hat{\beta}_i^s}}$, yielding the same Z-score value $\hat{Z}_i$. It is easy to see that we could have also perturbed Y (or any element of $\boldsymbol{U}$ for that matter) without changing $\hat{Z}_i$. Thus ordinary linear regression is invariant under shifting and scaling. Standardization of G_i to unit variance means that $\frac{1}{s}$ is the standard deviation of G_i. If the original GWAS sample should not be available to estimate this, reference data from projects such as 1000 Genomes [63], HapMap [77] or UK10K [78] can be used. The standard deviation of Y could similarly be estimated from publicly available data for many traits.

Though we have so far allowed covariates in the model, it is only possible if we have individual-level data. The availability of covariate information can not be assumed in summary statistics-based method development. Thus it is often assumed that Y was adjusted for covariates prior to association testing with genetic variants. In simple linear regression with SNP G_i, we can apply the least squares formula for variance 3.7 together with the maximum likelihood (ML) estimate 3.8 (assuming normal errors) for the residual variance to derive the estimated effect of standardized G_i on standardized Y:

$$\hat{\beta}_i^s = \hat{Z}_i\hat{\sigma}_{\hat{\beta}_i^s} = \hat{Z}_i\sqrt{\frac{1-\left(\hat{\beta}_i^s\right)^2}{n}} \implies \hat{\beta}_i^s = \operatorname{sign}\left(\hat{Z}_i\right)\sqrt{\frac{\hat{Z}_i^2}{n+\hat{Z}_i^2}}\,. \tag{3.14}$$

Thus we need not estimate standard deviations of variables to measure the change of Y in standard deviations per standard deviation change in G_i.

3.2.2. Binary outcome

The results above are derived for linear regression but binary traits are usually modelled via logistic regression using logit as a link function:

$$\operatorname{logit}(p) = \ln\frac{p}{1-p} = \beta_0 + G_i\beta_i\,, \tag{3.15}$$

such that $Y \mid G_i \sim \mathcal{B}(1,p)$. In this case, the least squares solution 3.5 does not apply, in fact logistic regression has no closed form solution. Note however that Y is a binary variable encoding for the presence of a disease. It is reasonable to assume that contracting the disease is not a simple on/off process. Instead, a continuous liability to the disease is assumed underneath and the disease is triggered if the liability exceeds a certain threshold. We can formalize this concept with a liability threshold model:

$$L = \beta_0 + G_i\beta_i + \varepsilon\,, \tag{3.16}$$

where L is a latent random variable for the liability and $\varepsilon \sim \text{Logistic}(0,1)$ with a cumulative distribution function $F_\varepsilon(x) = \frac{1}{1+\mathrm{e}^{-x}}$. Using an indicator variable, we can write $Y = \mathbb{1}_{L>0}$ and derive

$$\begin{aligned}
\mathrm{P}(Y=1) &= \mathrm{P}(\beta_0 + G_i\beta_i + \varepsilon > 0) \\
&= 1 - F_\varepsilon(-(\beta_0 + G_i\beta_i)) \\
&= \frac{1}{1+\mathrm{e}^{-(\beta_0+G_i\beta_i)}},
\end{aligned}$$

which corresponds to the logistic regression model 3.15 (through the same reasoning, it would correspond to probit regression if errors ε were normally distributed). It means we can treat parameter estimates in logistic regression as parameter estimates in the liability threshold model 3.16 [79] with the least squares solution 3.5; even normality of the estimator can be assumed to hold asymptotically (Subsection 5.4.1). Thus all summary statistics-based solutions can be applied on the effect estimates of both linear and logistic regression.

In the following, when dealing with parameter estimates from logistic regression in terms of logarithm of odds ratio (log-OR), we will simply assume to work on the liability scale, even though transformations between log-OR estimates of logistic regression and OLS-estimates of linear regression also exist [80]. Thus we will consider only linear regression models from now on.

4. ASSOCIATION BASED GENE PRIORITIZATION

> Each problem that I solved became a rule which served afterwards to solve other problems
>
> René Descartes

Computational analysis-based strategies of prioritizing genes for follow-up studies as candidates for disease intervention all depend on the concept of GWAS as the central building block. While GWAS on its own has fallen short in terms of providing mechanistic understanding of disease processes [46], its basic methodology has been extended to answer questions such as which genetic variants in the genome are functionally disease-relevant (fine-mapping), whether these variants simultaneously regulate gene expression traits (colocalization), and could we possibly approximate gene expression based on these variants to uncover putative disease-relevant genes in association studies with complex traits (TWAS). These questions do not facilitate answers in terms of causal gene-trait relationships, thus we will purposefully refrain from using formal causal language in this chapter. However, answers to these questions should at least get us closer to causal discoveries and we will explore that in the following sections.

4.1. Fine-mapping

Statistically significant SNPs from GWAS typically group together into larger genomic regions due to LD (Figure 3). A simple follow-up strategy for GWAS is to assume that each of these regions harbours a single causal variant—the one with the greatest $|\hat{Z}_i|$—and other variants are simply correlated with it. The causal variant could then be linked to a target gene with a TSS closest to it. This strategy, while seemingly naive, works surprisingly well at identifying true causal genes—at least in metabolite studies where experimental knowledge of functional relationships is abundant [81]. How well this strategy performs in studies of downstream complex traits where the association signals are generally much weaker is less clear and harder to validate. At least based on computational studies integrating eQTL and GWAS variants, the majority of trait-associated putative causal genes were not closest to GWAS loci [45].

It is also possible for a GWAS locus to harbour multiple causal variants. Let $\xi : m \times 1$ be a binary causal configuration vector where $\xi_i = 1$ indicates that the i-th SNP is causal. A comprehensive approach to estimating ξ would be to consider every possible combination of SNPs in a GWAS locus, train corresponding models and choose the one performing best based on some evaluation measure (e.g. Akaike criterium or coefficient of determination on hold-out data). This is computationally demanding and can even be infeasible if m is large since there are 2^m combinations of SNPs. Furthermore, ξ is thought to be sparse—an assumption

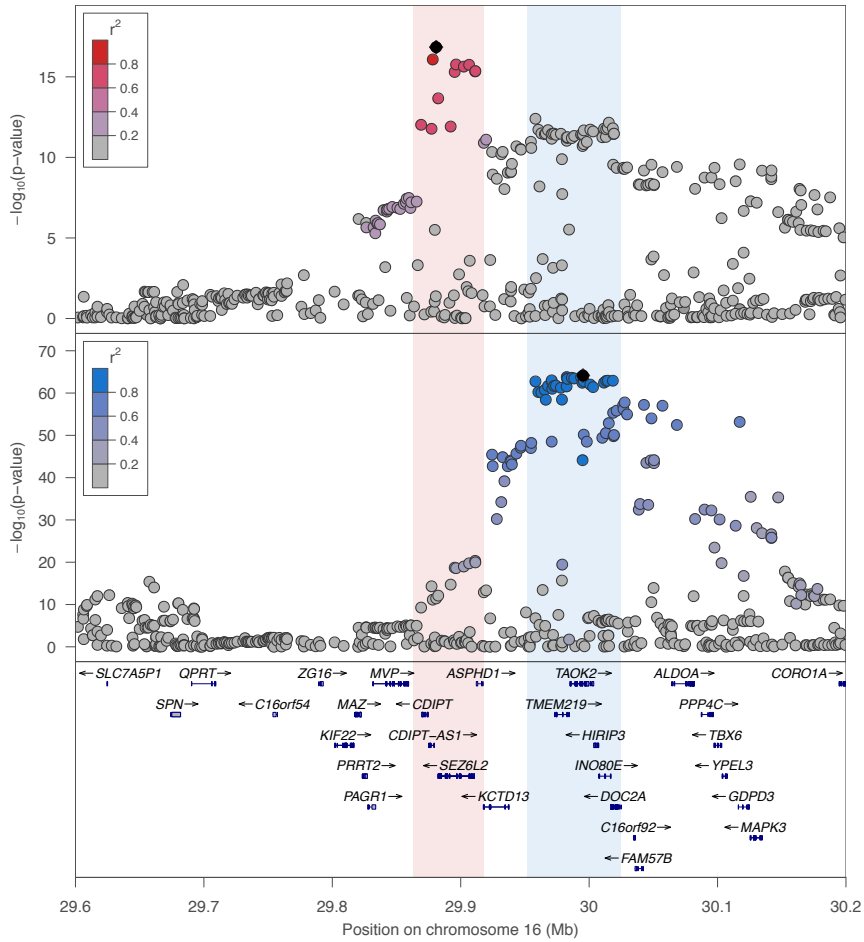


Figure 3: LocusZoom-derived [82] Manhattan plots depicting strengths of association of genetic variants with age at menarche (top) and INO80E expression (middle) in the 16p11.2 region (bottom). Each point corresponds to an association with a single SNP and is colored based on LD-correlation (r^2) with the lead SNP (black). The red and blue regions correspond to clusters of closely related GWAS and eQTL hits, respectively.

which general linear models do not necessarily account for [44]. Sparsity can be enforced by regularized regression but high LD among SNPs can result in a model which includes only non-causal SNPs [19].

4.1.1. Stepwise conditional analysis

The configuration vector ξ can be estimated by a forward selection heuristic in a stepwise application of the regression model 3.1—starting from the model with the most trait-associated SNP which is always assumed to be causal, other SNPs in the locus are selected based on their association P-values in the presence of previously selected SNPs. In particular, if we denote by $\boldsymbol{G}$ the set of all genotypes

G_i, then the set of selected SNPs at the i-th iteration is

$$\boldsymbol{G}_{(i)} = \begin{cases} \underset{G_j \in \boldsymbol{G}}{\operatorname{argmax}} |\hat{Z}_j|, & i = 1 \\ \boldsymbol{G}_{(i-1)} \cup \underset{G_j \in \boldsymbol{G} \setminus \boldsymbol{G}_{(i-1)}}{\operatorname{argmax}} |\hat{Z}_j|, & i > 1 \end{cases},$$

where Z-scores $\hat{Z}_j$ of SNPs G_j are the original GWAS Z-scores from models with a single SNP for $i = 1$ but for $i \geq 2$—by taking the liberty to treat the set $\boldsymbol{G}_{(i-1)}$ of random variables as a random vector—are calculated from the model below:

$$Y = G_i \beta_i + \boldsymbol{G}'_{(i-1)} \beta_{(i-1)} + \varepsilon. \tag{4.1}$$

The procedure is stopped once the maximal $|\hat{Z}_j|$ in an iteration does not exceed a pre-specified threshold. As a result, we have $\hat{\xi}_i = \mathbb{1}_{G_i \in \boldsymbol{G}_*}$, where $\boldsymbol{G}_*$ is the final set of variables.

Stepwise models can also be implemented for summary statistics. Consider a random vector of centered and scaled genetic variants $\boldsymbol{G} = (G_1, G_2, \ldots, G_m)$ and let $G : n \times m$ be the corresponding design matrix. The OLS solution 3.5 enables to express (joint) effect estimates $\hat{\beta}_J$ from the multiple regression model $Y = \boldsymbol{G}\beta_J + \varepsilon$ in terms of (marginal) GWAS effect estimates $\hat{\beta}_i$ from simple linear regression models $Y = G_i \beta_i + \varepsilon$ as follows [83]:

$$\hat{\beta}_J = \Sigma^{-1} \hat{\beta}_M, \tag{4.2}$$

$$\operatorname{Var}(\hat{\beta}_J) = \sigma^2_{\varepsilon_J} n^{-1} \Sigma^{-1}, \tag{4.3}$$

where $\hat{\beta}_M = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_m)$ and $\Sigma = G'G$ is the correlation matrix of SNPs. The latter can be estimated from an external reference panel if the original sample should be unavailable. Since OLS residuals are uncorrelated with regressors by design, the ML estimate for the error variance 3.8 in the joint model is

$$\begin{aligned} \hat{\sigma}^2_{\varepsilon_J} &= \frac{1}{n}(y - G\hat{\beta}_J)'(y - G\hat{\beta}_J) \\ &= \frac{1}{n}(y'y - \hat{\beta}'_J (G'G)\hat{\beta}_J) \\ &= 1 - \hat{\beta}'_J \Sigma \hat{\beta}_J \,, \end{aligned} \tag{4.4}$$

Thus we can calculate Z-scores in the joint model using only quantities which we know or have access to. Summary statistics-based stepwise models as described above have been implemented in GCTA software [84].

Stepwise models belong to the class of greedy algorithms which work by making a locally optimal choice at each iteration. Such heuristics are straightforward to implement and can sometimes work but may miss the global optimum. Further, it is not clear what should be the Z-score threshold to keep adding SNPs in the model. Using a genome-wide threshold may result in a lack of power to detect secondary signals and true causal variants but relaxing the threshold can induce unwanted side-effects in terms of inclusion of non-causal SNPs [19].

4.1.2. Bayesian fine-mapping

Specifically designed for fine-mapping purposes are Bayesian variable selection methods which work by estimating the posterior probability $p(\xi \mid y, G)$ of any causal configuration vector ξ, given that we have observed data (y, G). The sparsity of the true ξ can be enforced via prior distribution $p(\xi)$ and the probabilistic outcome leads to a natural interpretation of the solution which is advantageous over competing methods. Making use of the Bayes formula,

$$p(\xi \mid y, G) = \frac{p(y \mid \xi, G)p(\xi \mid G)}{p(y \mid G)} \propto p(y \mid \xi, G)p(\xi), \tag{4.5}$$

where proportionality holds because $p(y \mid G)$ does not depend on ξ. The prior $p(\xi)$ can simply assume each SNP equally likely to be causal and be defined in terms of the probability to observe any total number of causal SNPs [43, 44] but may also incorporate additional information in terms of functional or other annotations to prioritize certain causal variants [85,86]. We do not know the exact form of the likelihood $p(y \mid \xi, G)$ but we can marginalize over parameters β by assuming normality $y \mid (\beta, G) \sim \mathcal{N}(G\beta, \sigma_\varepsilon^2 I)$ in the linear regression model 3.4; and define $\beta \mid (\xi, G) \sim \mathcal{N}(0, \Lambda_\xi)$, where Λ_ξ is a diagonal matrix where the i-th element on the diagonal is some pre-specified positive value if $\xi_i = 1$, otherwise a small constant close to zero such that Λ_ξ would be invertible. Following chapter 2.3.3 in Bishop [87],

$$\begin{aligned} p(y \mid \xi, G) &= \int p(y \mid \beta, G)p(\beta \mid \xi, G)d\beta \\ &\propto \exp\left(-\frac{1}{2}y'(\sigma_\varepsilon^2 I + G\Lambda_\xi G')^{-1}y\right), \end{aligned} \tag{4.6}$$

which corresponds to the probability density function of $\mathcal{N}(0, \sigma_\varepsilon^2 I + G\Lambda_\xi G')$.

In order to express likelihood 4.6 in terms of summary statistics, we can use the Woodbury matrix identity to rewrite the matrix inverse and then make use of equations 4.4 and 3.5 to express all terms including y as some functions of $\hat{\beta}_J$. Doing so, we obtain

$$\begin{aligned} p(y \mid \xi, G) &\propto \exp\left(-\frac{1}{2}\hat{\beta}_J'\left(V^{-1} + V^{-1}(V^{-1} + \Lambda_\xi^{-1})^{-1}V^{-1}\right)\hat{\beta}_J\right) \\ &= \exp\left(-\frac{1}{2}\hat{\beta}_J'(V + \Lambda_\xi)^{-1}\hat{\beta}_J\right), \end{aligned} \tag{4.7}$$

where the equality follows after once again applying the Woodbury matrix identity and V is a shorthand for $\mathrm{Var}(\hat{\beta}_J)$ from equation 4.3. The exponent in 4.7 corresponds to the probability density function of $\mathcal{N}(\hat{\beta}_J; 0, V + \Lambda_\xi)$ and can be calculated based on summary statistics alone. Some examples of software where such summary statistics-based fine-mapping has been implemented include CAVIAR [43], FINEMAP [44], and PAINTOR [86].

Knowing the posterior probability of any causal configuration vector makes it possible to estimate the true ξ as the one with maximal probability. Bayesian fine-mapping studies usually go a step further and report sets of genetic variants that contain all the causal variants with some probability. These so-called credible sets are used to prioritize causal SNPs in follow-up studies.

4.2. Colocalization

Fine-mapping improves the identification of causal SNPs in genomic loci but these variants could be linked to genes only based on physical proximity to TSS—unless there was additional data. Since most GWAS-associated loci are found in gene regulatory regions (Subsection 2.2.2), it is reasonable to assume that gene expression plays the role of a mediator between the causal variants and complex traits. Thus integrating knowledge from transcriptomics data should enhance our chances of teasing out functional genes as complex trait-associated SNPs that are also associated with gene expression levels (such SNPs are called eQTLs) are likely to manifest their effects on respective traits through regulation of corresponding genes. However, even if significant loci in the eQTL and GWAS studies overlap, LD patterns make it difficult to interpret whether the same causal variant is responsible for both signals (Figure 3). Colocalization analyses have been devised to answer precisely this question.

Let $\xi^{(c)} : m \times 1$ be the causal configuration vector of a GWAS locus for the complex trait of interest. Let $\xi^{(e)} : m \times 1$ be the causal configuration vector of the same locus in the eQTL study of some gene. Formally, if there is a shared causal variant in any of the $i \in \mathcal{I}_m$ positions of the locus such that $\xi_i^{(c)} = \xi_i^{(e)} = 1$, then we have support for a causal interpretation between the gene and complex trait.

4.2.1. Bayesian colocalization

Colocalization can be thought of as an extension of Bayesian fine-mapping to multiple datasets, thus the underlying theory is very similar and based on identifying the posterior probability of observing a shared causal variant, given both the eQTL and GWAS data. Since eQTL and GWAS studies are usually performed on independent samples, we can use the independence property of the joint probability and write

$$p\left(\xi_i^{(c)} = 1, \xi_i^{(e)} = 1 \mid D^{(c)}, D^{(e)}\right) = p\left(\xi_i^{(c)} = 1 \mid D^{(c)}\right) p\left(\xi_i^{(e)} = 1 \mid D^{(e)}\right), \tag{4.8}$$

where $D^{(c)} = (y, G)^{(c)}$ and $D^{(e)} = (y, G)^{(e)}$ denote the observed data in GWAS and eQTL studies, respectively. The probability in 4.8 is called colocalization posterior probability (CLPP). There can be many causal configuration vectors with a 1 in the i-th position, thus we can write

$$p(\xi_i = 1 \mid D) = \sum_{\xi : \xi_i = 1} p(\xi \mid y, G). \tag{4.9}$$

Note that the summands in the equation above are exactly the same as defined in equation 4.5, thus we can reuse all the theory developed in Subsection 4.1.2 to calculate the CLPP for any position i in equation 4.8. The software eCAVIAR [49], a direct generalization of CAVIAR, does exactly that. In this case, the gene is classified as functional for the complex trait and targeted for additional studies if any of the variants in the locus achieve a CLPP value greater than some pre-specified threshold.

We can simplify the analysis above if we assume at most one causal variant in the eQTL data and at most one causal variant in the GWAS data. Then the sum in equation 4.9 would have exactly one term and there could be no causal configurations ξ which could satisfy both $\xi_i = 1$ and $\xi_j = 1$ if $i \neq j$ (i.e. these events would be mutually exclusive). This would enable to calculate the posterior probability of a shared causal variant in any position of the locus, deciding for a shared causal variant if the sum

$$\sum_{\xi:\sum_{i=1}^{m}\xi_i=1} p\left(\xi \mid D^{(c)}\right) p\left(\xi \mid D^{(e)}\right)$$

exceeds some pre-specified threshold. The simplified approach is implemented in a popular tool called COLOC [48].

4.2.2. Non-Bayesian colocalization and ties to causality

Colocalization methods do not assign a causal effect size nor provide a formal test of causality between two traits. By definition, a shared causal variant between the traits can simply exhibit horizontal pleiotropy or happen as a result of reverse causation (e.g. the complex trait might have a causal effect on the gene expression trait, not vice versa as we might expect). However, there is a useful non-Bayesian alternative to colocalization analysis—based on exploiting the homogeneity of effect estimates in a locus with a single shared causal variant—which does more in terms of causality.

Consider a locus with a single causal variant G_*, shared between traits, and let G_i be a distinct variant with LD-correlation $r_{i*} > 0$. It follows that in an unconfounded multiple regression model with both of these variants, the true effect of G_i is 0 (see Subsection 5.2.1 and Subsection 5.3.1 below). Unbiasedness 3.6 of OLS effect estimates enables us to exploit relationship 4.2 between joint and marginal effects—without loss of generality, assume both variants are centered and standardized to unit variance—to express the true marginal effect of G_i in terms of the true marginal effect of G_* as $\beta_i = \beta_* r_{i*}$. Crucially, this relation holds for the summary statistics of any trait, thus the ratio of GWAS and eQTL true marginal effects of different genetic variants G_i are equal (or homogeneous) due to LD-correlation canceling out, $b_i = \frac{\beta_i^{(c)}}{\beta_i^{(e)}} = \frac{\beta_*^{(c)} r_{i*}}{\beta_*^{(e)} r_{i*}} = b_*$. Since $\hat{b}_i$ are approximatey

normal (see 5.13), we can approximate as follows [88]:

$$\hat{d}_i = \underbrace{\hat{b}_i}_{\substack{\| \\ \frac{\hat{\beta}_i^{(c)}}{\hat{\beta}_i^{(e)}}}} - \underbrace{\hat{b}_*}_{\substack{\| \\ \frac{\hat{\beta}_*^{(c)}}{\hat{\beta}_*^{(e)}}}}, \qquad \hat{d}_i \sim \mathcal{N}\left(0, \underbrace{\mathrm{Var}(\hat{d}_i)}_{\substack{\| \\ \mathrm{Var}(\hat{b}_i) + \mathrm{Var}(\hat{b}_*) - 2\mathrm{Cov}(\hat{b}_i, \hat{b}_*)}}\right), \tag{4.10}$$

where the components of $\mathrm{Var}(\hat{d}_i)$ can be derived using the Delta method (see Appendix A for the derivation).

Testing for colocalization is equivalent to testing for homogeneity in estimates $\hat{d}_i$ of genetic variants in the locus of interest. Recall that we encountered a similar problem in Subsection 3.1.2 when testing for the homogeneity of effect estimates in meta-analysis. Back then we used the Cochran's Q statistic 3.13 for this purpose but it assumed independent estimates while $\mathrm{Cov}(\hat{d}_i, \hat{d}_j)$—each component of it is of the form $\mathrm{Cov}(\hat{b}_\cdot, \hat{b}_\cdot)$ and can thus be derived using the Delta method (see Appendix A for details)—can be non-zero. We can nevertheless follow the same procedure by standardizing $\hat{d}_i$ to Z-scores and then taking the squares before summing together. Doing so results in a test statistic for the heterogeneity in dependent instruments (HEIDI, where instrument means genetic variant)

$$T_{HEIDI} = \sum_{i \in \mathcal{I}_m \setminus \{*\}} \frac{\hat{d}_i^2}{\mathrm{Var}(\hat{d}_i)}, \tag{4.11}$$

which follows a generalized chi-squared distribution and does not have a closed form, but can be approximated by numerical methods [88].

Similarly to Bayesian colocalization methods, using the HEIDI test statistic 4.11 to test for colocalization provides no information about the causal effect size nor even the direction of it. However, in constructing T_{HEIDI} we relied heavily on quantities $\hat{b}_i$. It turns out that under some assumptions, $\hat{b}_i$ is actually an estimate for a causal effect b_i between the two traits. This is a useful observation and we will treat this in great detail in later chapters.

4.3. Transcriptome-wide association studies

Remember that genome-wide analyses go together with huge multiple testing burdens due to the sheer number of genetic variants. Controlling the type 1 error (T1E) rate reduces statistical power proportionally to the number of performed tests, making it hard to detect weaker signals as a consequence. This has downstream effects on colocalization analyses since testing for shared causal variants makes sense only if significant results were found in both GWAS and eQTL studies. Reducing the multiple testing burden should thus have obvious benefits; the question remains how to do so in a statistically informed manner. TWAS attempts it by first aggregating genetic information on the gene level before proceeding with the association analysis [89]. Since there are two orders of magnitude fewer

effective number of independent genes than SNPs, this arrangement looks promising in terms of identifying more functionally relevant genes.

Consider observations $y : n \times 1$ of an outcome random variable Y, $x : n \times 1$ of a gene expression random variable X, and $G : n \times m$ of a random vector $\boldsymbol{G} : m \times 1$ of genetic variants. Without loss of generality, let us assume that all variables are centered and scaled to unit variance (if not, we can always do that). A TWAS can effectively be thought of as a two-stage regression analysis. In the first stage, gene expression levels are approximated by a linear combination of genetic variants:

$$\hat{x} = \sum_{i=1}^{m} w_i G_{\cdot i} = Gw, \tag{4.12}$$

where $w = (w_1, w_2, \ldots, w_m)'$ is a weight vector representing eQTL effect sizes from any form of (penalized) regression analysis. In the second stage of TWAS, these approximations are tested for an association with the outcome in a linear regression model, yielding OLS estimates 3.5 and 3.7 for the effect size and variance, respectively:

$$\hat{\beta}_{\hat{X}Y} = (\hat{x}'\hat{x})^{-1}\hat{x}'y, \tag{4.13}$$

$$\mathrm{Var}\big(\hat{\beta}_{\hat{X}Y}\big) = \sigma_\varepsilon^2(\hat{x}'\hat{x})^{-1}. \tag{4.14}$$

The Wald test can then be used to test for the significance of association between the gene expression trait X and outcome Y.

Note that TWAS as presented above requires individual-level data (e.g. software PrediXcan [50]). However, it is straightforward to extend the procedure to work on summary statistics. Let $\hat{\beta}_M = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_m)'$ be the vector of (marginal) GWAS effect sizes of individual genetic variants on the outcome and let $\Sigma = G'G$ denote the LD-correlation matrix of respective variants. Plugging equation 4.12 into equations 4.13 and 4.14 leads to a derivation of these formulae based on those quantities (S-PrediXcan [52]):

$$\hat{\beta}_{\hat{X}Y} = (\hat{x}'\hat{x})^{-1}\hat{x}'y = \frac{w'G'y}{w'G'Gw} = \frac{w'\hat{\beta}_M}{w'\Sigma w}, \tag{4.15}$$

$$\mathrm{Var}(\hat{\beta}_{\hat{X}Y}) = \sigma_\varepsilon^2(\hat{x}'\hat{x})^{-1} = \frac{\sigma_\varepsilon^2}{w'G'Gw} = \frac{\sigma_\varepsilon^2}{nw'\Sigma w}. \tag{4.16}$$

In order to complete the derivation in terms of the error variance σ_ε^2, note first that the second stage regression in TWAS can be written as a multiple regression model $Y = \mathbf{G}'w\beta_{\hat{X}Y} + \varepsilon$. This observation enables us to make use of equation 4.4 to derive the residual variance as follows:

$$\hat{\sigma}_\varepsilon^2 = 1 - (w\hat{\beta}_{\hat{X}Y})'\Sigma(w\hat{\beta}_{\hat{X}Y}) = 1 - \frac{(w'\hat{\beta}_M)^2}{w'\Sigma w}.$$

We could also take $\sigma_\varepsilon^2 \approx 1$ if genetic variants (SNPs) described only a tiny portion of gene expression variability. Such a simplification allows to test for the significance of TWAS effects (i.e. whether the null hypothesis $H_0 : \beta_{\hat{X}Y} = 0$ holds)

through a linear combination of GWAS Z-scores $\hat{Z}_M = (\hat{Z}_1, \hat{Z}_2, \ldots, \hat{Z}_m)$ as follows: $\hat{Z}_{\hat{X}Y} = \frac{w'\hat{Z}_M}{\sqrt{w'\Sigma w}} \sim \mathcal{N}(0,1)$ [51]. This result is not only convenient but also works very much like summary statistics imputation methods [90, 91]. After all, we are also trying to exploit available information to impute related quantities.

4.3.1. TWAS for implicating causal genes

Note that gene expression approximations 4.12 are simply genetically predicted values of X, or in other words—genetic risk scores. Since TWAS investigates the trait-relevance of genetically informed expression, there might be a desire to consider significant associations as causal relationships, exactly like in GWAS. However, while the design of TWAS indeed makes reverse causality unlikely—identifying eQTLs in the first step biases away from this—causality can actually be claimed only when genetic variants satisfy certain assumptions, e.g. G_i should have no horizontal pleiotropy in terms of X and Y [92]. Gene co-regulation as a form of pleiotropy, such as due to sharing or having LD-correlated eQTLs with the true causal gene, might result in non-causal genes falsely showing up as TWAS hits. To correct for this bias, gene-level fine-mapping—based on the same theory as presented in Section 4.1 but focusing on the PIPs and credible sets of genes, not SNPs—has recently been proposed; however, more work is said to be needed [53]. Furthermore, the second stage regression in TWAS does not account for the uncertainty in predictions $\hat{x}$. Though not without its merits, TWAS can thus be considered an *ad hoc* method for identifying causality, or even invalid (see Subsection 6.2.3 for more details).

For reasons brought above, causal reasoning based on TWAS can be dangerous. Somewhat paradoxically however, TWAS greatly resembles MR, a method of instrumental variables which is specifically designed for identifying causal relationships in genetics. Indeed, if w_i in 4.12 were calculated from a multiple regression model, the TWAS estimator 4.13 would be equal to a two-stage least squares estimator 5.17, the go-to estimator in MR for implying causality. In order to understand what is going on, we will properly treat the concept of causality in the subsequent Chapter 5, followed by an extensive focus on MR in Chapter 6.

5. CAUSAL INFERENCE

> We do not know a truth without knowing its cause
>
> Aristotle

We have thus far introduced several approaches through which causal reasoning is regularly attempted in genetics research. While the methods we have covered have been used to implicate gene targets in the past [49, 53, 93–95], they are not formal tests of causality between gene expression and complex traits. In order to do better, we must first define what we mean by causality.

5.1. Causal relationships

Consider a biological system represented by a set of random variables $\mathcal{V}$ and the following cause-effect relationships between them:

$$\begin{aligned} &V = f_V\left(\text{parents}(V), \varepsilon_V\right) \text{ for all } V \in \mathcal{V}, \\ &\{\varepsilon_V : V \in \mathcal{V}\} \text{ are mutually independent,} \end{aligned} \tag{5.1}$$

where $\text{parents}(V)$ and ε_V respectively refer to all the causes of V (also called the effect) included and not included in $\mathcal{V}$, and f_V denote for arbitrary functions. The mutual independence of ε_V ensures that $\mathcal{V}$ entails all the relevant information for characterising the relationships between the variables in the system. While feedback loops between variables are possible in reality, we assume acyclicity to facilitate statistical inference and it makes sense if we treat each variable in reference to time—loops could not occur in this case and the same phenomenon at a later time point would simply have to be denoted differently.

Cause-effect (causal) relationships imply a direction of effect whereby one event (the cause) precedes the other (the effect) [40]. For example, a disease, once severe enough, leads to a symptom; by the central dogma of molecular biology, mRNA molecules injected into a human organism will be translated to more proteins but added proteins will not be transformed into respective mRNAs (Subsection 2.1.1). Put simply, a causal relationship cannot be reversed. This is a requirement mathematical equations do not accommodate for. We could escape this restriction by using a different notation to describe directionality between variables (e.g. assigment symbols ":=" or "$\leftarrow$" instead of the equality "="). Instead, a common practice to facilitate causal language is to complement mathematical equations with graphical models.

5.1.1. Directed acyclic graphs as causal models

Let us represent the presence of a cause-effect relationship between any two variables in $\mathcal{V}$ as a directed edge, and let $\mathcal{E}$ be the set of all such edges. The biological

system 5.1 can then be depicted as a directed acyclic graph (DAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the inclusion of ε_V is made redundant due to the assumption that variables in $\mathcal{V}$ can only be related with each other through directed edges in $\mathcal{E}$. The directed edges are represented as arrows in the DAG—the parents of $V \in \mathcal{V}$ have an outgoing arrow to V while the latter has incoming arrows from its parents.

The cause-effect relationships in $\mathcal{E}$ are also called parent-child relationships. In similar (kinship) fashion, variables up- and downstream of V are called its ancestors and descendants, respectively. A causal relationship from $X \in \mathcal{V}$ to $Y \in \mathcal{V}$ implies a sequence of arrows pointing from X to Y in the DAG. The relationship is direct if there is a directed edge $X \to Y$ in $\mathcal{E}$ (i.e. X is a parent and Y is a child), otherwise it is mediated by other variables and thus indirect (i.e. X is an ancestor and Y is a descendant). In effect, a DAG is a graphical representation of causal relationships between random variables in a causal system (an example DAG is brought in Figure 4a). When referring to a DAG in the following, we mean the underlying causal system it represents.

5.1.2. Intervention in the causal system

To decipher causal relationships in the (biological) system represented by the DAG, we need to be able to detect changes to the system due to external events. This comes down to evaluating the probability to observe any particular state of variables of the DAG. Since the value of every random variable $V_i \in \mathcal{V}$ depends only on its parents' values, the probability distribution of V_i conditional on the random vector $\boldsymbol{V}_{(-i)}$ of all other variables in $\mathcal{V}$ satisfies local Markov property [96]:

$$p\left(V_i = v_i \mid \boldsymbol{V}_{(-i)} = \boldsymbol{v}_{(-i)}\right) = p(V_i = v_i \mid \text{parents}(V_i)).$$

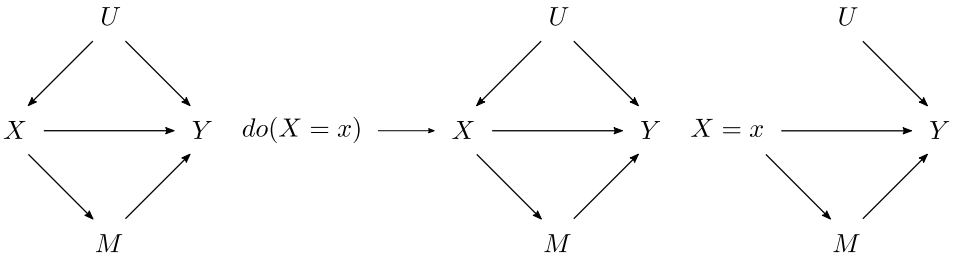
The joint probability distribution of all variables $\left(V_i \in \mathcal{V} : i \in \mathcal{I}_{|\mathcal{V}|}\right)$ encoded by $\mathcal{G}$, by applying the chain rule of probability, is thus

$$p(\mathcal{G}) := p\left(V_1 = v_1, V_2 = v_2, \ldots, V_{|\mathcal{V}|} = v_{|\mathcal{V}|}\right) = \prod_{i=1}^{|\mathcal{V}|} p\left(V_i = v_i \mid \text{parents}(V_i)\right). \quad (5.2)$$

The probability 5.2 is also called pre-intervention probability [40]. An external intervention on any of the variables of the DAG (e.g. administration of a drug) perturbs the DAG and thus affects the rule to calculate the joint probability of the variables encoded by it.

We define an intervention on a random variable $X \in \mathcal{V}$ as an act which overrides the effect of all factors affecting the variable and imposes on it a value x, such that $X = f_X(\text{parents}(X), \varepsilon_X)$ becomes simply $X = x$. Following Pearl [97], we denote the intervention by $do(X = x)$. Since the value of X is artificially fixed, all its incoming edges (parents) in the corresponding DAG $\mathcal{G}$ are removed (Figure 4b). The probability 5.2 becomes

$$p\left(\mathcal{G} \mid do(X = x)\right) = \prod_{\substack{V_i \in \mathcal{V} \\ V_i \neq X}} p\left(V_i = v_i \mid \text{parents}(V_i)\right) = \frac{p(\mathcal{G})}{p(X = x \mid \text{parents}(X))} \quad (5.3)$$



(a) (b) (c)

Figure 4: (a) Causal graph depicting a direct causal effect between X and Y, a mediator M, and a confounder U. (b) The same graph depicting an external intervention of setting X to a value x (i.e. $do(X = x)$); the variable X becomes a collider between the intervention and its parent U. (c) Modified causal graph where the intervention $do(X = x)$ has forced the value of X to x and thus overridden the effect of U on X; the post-intervention probability of the variables is to be calculated from the modified graph.

and is called post-intervention probability [40].

It follows immediately from equation 5.3 that should variable X have no parents, the post-intervention probability $p(\mathcal{G} \mid do(X = x))$ would be exactly equal to the conditional probability $p(\mathcal{G} \mid X = x)$. In general, this need not be the case. To grasp the difference between the two quantities, consider that the latter examines the probability to observe a particular variable state among instances of $X = x$ while the former examines it after imposing a universal rule $X = x$ over the entire sample space. In an example with two binary random variables X and Y coding for a treatment and a disease, respectively, $p(Y = \text{yes} \mid X = \text{yes})$ represents the probability to observe a disease among those who got the treatment but $p(Y = \text{yes} \mid do(X = \text{yes}))$ represents the probability to observe a disease had everyone got the treatment. It makes sense that the two quantities are equal only if treatment assignment was not systematically influenced by external factors. Conditional probabilities are associative in nature; to obtain causal effects, we need to work with post-intervention probabilities.

5.1.3. Intervention's effect on the outcome—the causal effect

In general, we are not interested in possible changes to the state of variables of the entire causal system, rather we are interested in quantifying the intervention's effect—the causal effect—to a particular variable. To establish causality between any random variables X and Y, we would need to determine whether different interventions on X would result in perturbations in Y:

$$p(Y = y \mid do(X = x_1)) \neq p(Y = y \mid do(X = x_2)) \tag{5.4}$$

for some values x_1, x_2 of X and y of Y. Of course, in practice it is not possible to enforce the value of X over the entire sample space and observe what hap-

pens. The *do*-operator is thus a fictional concept. Fortunately, causality can still be determined in some cases and under some assumptions. We would just need to express $p(Y = y \mid do(X = x))$ in terms of quantities we can calculate from observational data, such as conditional probabilities of the form $p(Y = y \mid X = x)$.

Note that we can express the post-intervention probability of Y by marginalizing over any set of variables $\boldsymbol{V} \subseteq \mathcal{V} \setminus \{X, Y\}$ of the DAG:

$$p(Y = y \mid do(X = x)) = \int p(Y = y, \boldsymbol{V} = \boldsymbol{v} \mid do(X = x)) d\boldsymbol{v}. \tag{5.5}$$

We can get rid of the *do*-operator on the right-hand side of equation 5.5 by taking $\boldsymbol{V} = \mathcal{V} \setminus \{X, Y\}$ [40]. In this case, the probability under the integral is simply $p(\mathcal{G} \mid do(X = x))$ which was expressed in equation 5.3 using quantities we can compute from observational data. Formally, for $\boldsymbol{V} = \mathcal{V} \setminus \{X, Y\}$,

$$\begin{aligned} p(Y = y \mid do(X = x)) &= \int \frac{p(Y = y, X = x, \boldsymbol{V} = \boldsymbol{v})}{p(X = x \mid \text{parents}(X))} d\boldsymbol{v} \\ &= \int \frac{p(Y = y, X = x \mid \boldsymbol{V} = \boldsymbol{v})}{p(X = x \mid \text{parents}(X))} p(\boldsymbol{V} = \boldsymbol{v}) d\boldsymbol{v} \\ &= \mathrm{E}_{\boldsymbol{V}} \left[\frac{p(Y = y, X = x \mid \boldsymbol{V} = \boldsymbol{v})}{p(X = x \mid \text{parents}(X))} \right]. \end{aligned} \tag{5.6}$$

As usual, the expected value $\mathrm{E}_{\boldsymbol{V}}$ (over variables $\boldsymbol{V}$) in equation 5.6 reduces to a sum over the probability space in case of discrete random variables and even in case of continuous variables can always be estimated using statistical techniques by enforcing some distributional constraints.

Equation 5.6 is remarkable in the sense that it allows us to estimate the distribution of the complex trait Y on the assumption that we had intervened on X by artificially setting its value to some x, without actually making said intervention. Unfortunately, calculating the post-intervention probability of Y using equation 5.6 requires not only complete understanding of the causal relationships between variables (knowledge of the graph structure), but also that we have measured all said variables. These requirements are never met in biological settings. Luckily, it turns out the causal effect is identifiable even when only a sufficient subset of variables $\boldsymbol{V}_s \subseteq \mathcal{V} \setminus \{X, Y\}$ is known and measured.

5.2. Identifiability of the causal effect

Consider again equation 5.5 for calculating the probability $p(Y = y \mid do(X = x))$. In this equation, the joint probability under the integral can be expressed via conditional probability to obtain

$$p(Y = y \mid do(X = x)) = \int p(Y = y \mid \boldsymbol{V} = \boldsymbol{v}, do(X = x)) p(\boldsymbol{V} = \boldsymbol{v} \mid do(X = x)) d\boldsymbol{v}.$$

Note that we can get rid of both intervention terms $do(X = x)$ under the integral whenever $\boldsymbol{V}$ is such that it is not affected by intervention on X and, given knowledge about it, makes Y also not affected. Any set $\boldsymbol{V}_s \subseteq \mathcal{V} \setminus \{X, Y\}$ satisfying these two criteria is called sufficient and enables us to express the post-intervention probability using familiar quantities [98]:

$$\begin{aligned} p(Y = y \mid do(X = x)) &= \int p(Y = y \mid \boldsymbol{V}_s = \boldsymbol{v}_s, X = x) p(\boldsymbol{V}_s = \boldsymbol{v}_s) d\boldsymbol{v}_s \\ &= \mathrm{E}_{\boldsymbol{V}_s}\big[p(Y = y \mid \boldsymbol{V}_s = \boldsymbol{v}_s, X = x)\big]. \end{aligned} \tag{5.7}$$

A pertinent question is how to find a sufficient set $\boldsymbol{V}_s$ which would allow causal analysis on observational data. To answer this question, we will examine how to formulate the criteria it has to satisfy based on the graphical model (the DAG).

5.2.1. D-separation

Let us define a path between variables $X, Y \in \mathcal{V}$ in a DAG as any sequence of distinct variables $(V_1, V_2, \ldots, V_p)$ such that $V_1 = X$, $V_p = Y$, and V_{i+1} is either a parent or a child of V_i (i.e. we do not care about the directionality of edges in a path, unless we refer to a causal path where every V_{i+1} must be a child of V_i). Notice that any three consecutive variables V_i, V_{i+1}, V_{i+2} in a path can be related in the following three ways (see also Figure 4):

(a) $V_i \to V_{i+1} \to V_{i+2}$ constitutes a chain. Since V_{i+2} is a function of V_{i+1} which in turn is a function of V_i, it is clear that $V_i \not\perp\!\!\!\perp V_{i+2}$. However, the Markov property of the DAG [96] ensures that given knowledge about the value of V_{i+1}, the variable V_i provides no extra information, leading to conditional independence $V_i \perp\!\!\!\perp V_{i+2} \mid V_{i+1}$.

(b) $V_i \leftarrow V_{i+1} \to V_{i+2}$ constitutes a fork where the variable V_{i+1} in the middle is called a confounder. Since V_i and V_{i+2} are both functions of V_{i+1}, it is clear that $V_i \not\perp\!\!\!\perp V_{i+2}$. However, knowledge about V_{i+1} breakes this relationship, resulting in conditional independence $V_i \perp\!\!\!\perp V_{i+2} \mid V_{i+1}$.

(c) $V_i \to V_{i+1} \leftarrow V_{i+2}$ constitutes an inverted fork where the variable V_{i+1} in the middle is called a collider. Since neither V_i nor V_{i+2} depend on common ancestors, it is clear that $V_i \perp\!\!\!\perp V_{i+2}$. However, knowledge about the common child V_{i+1} (or any descendant of it) can be used to deduce some information about the parents, thus $V_i \not\perp\!\!\!\perp V_{i+2} \mid V_{i+1}$.

We say that a path between X and Y is blocked if there are colliders on this path. Blocked paths can be unblocked by conditioning on any of the colliders or their descendants (point (c) above). Similarly, unblocked paths can be blocked by conditioning on confounders (point (b)) or middle variables in chains (point (a)). Importantly, conditional independence between variables X and Y in a graph $\mathcal{G}$ requires—provided knowledge about some set of variables $\boldsymbol{V}$ which is allowed to be empty—that all paths between them are blocked. In that case, we say X and Y are directionally separated or d-separated [99].

Using the logic laid out above, sufficient sets $\boldsymbol{V}_s \subseteq \mathcal{V} \setminus \{X, Y\}$ for calculating the post-intervention probability $p(Y = y \mid do(X = x))$ in equation 5.7 can be constructed based on the graphical model by following a so-called back-door criterion [98]:

(i) No $V \in \boldsymbol{V}_s$ is a descendant of X.

(ii) The elements of $\boldsymbol{V}_s$ block all back-door paths from X to Y, i.e. those with an arrow pointing to X.

The first rule ensures that all paths between X and Y representing true causation (composed of chains) are unblocked while the second rule ensures that those paths representing spurious relationships due to confounding (containing forks) are blocked. If a sufficient set satisfying these rules can be constructed, then the causal effect between X and Y can be identified [98]. Moreover, we will see in the next section that we can use regression to estimate the causal effect.

5.3. Assuming linear causal effects

The causal theory covered above did not enforce any distributional assumptions on the random variables in the DAG. Equation 5.7 represents thus a general solution for calculating the causal effect. However, in gene expression studies we are willing to assume linearity between variables. This makes testing for causality 5.4 equivalent to testing for the difference in average causal effects of X on Y by a unit increase in X:

$$\mathrm{E}(Y \mid do(X = x)) - \mathrm{E}(Y \mid do(X = x - 1)). \tag{5.8}$$

Thus it is useful to translate causal inference into regression framework.

Linearity in a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ means that the value of each variable is determined by a linear combination of its parents' values. In other words, we can annotate the edges between variables $V_i, V_j \in \mathcal{V}$ as $V_i \xrightarrow{b_{V_i \to V_j}} V_j \in \mathcal{E}$ and write

$$V_j = b_{V_j} + \sum_{V_i \in \mathrm{parents}(V_j)} V_i \cdot b_{V_i \to V_j} + \varepsilon_{V_j},$$

where b_{V_j} is the intercept and ε_{V_j} denotes for all factors affecting V_j other than its parents in the DAG. That is, $b_{V_i \to V_j}$ is the direct causal effect of variable V_i on V_j and corresponds to a change in V_j per unit increase of V_i. In this work, we are interested in the total effect of an intervention $do(X)$ on an outcome Y, thus we also need to consider indirect effects through mediators. It is easy to see that the total causal effect b_{XY} of X on Y is the sum of products of direct effects in causal paths between X and Y:

$$b_{XY} = \sum_{\substack{\text{causal path} \\ \text{between } X \text{ and } Y}} \prod_{\substack{\text{edge } V_i \to V_j \\ \text{in the path}}} b_{V_i \to V_j}. \tag{5.9}$$

We can thus write [40]

$$\mathrm{E}(Y \mid do(X = x)) = \mathrm{E}(b_{XY}X + \varepsilon \mid do(X = x)) = b_{XY}x + \mathrm{E}(\varepsilon),$$

where ε covers the effects of all parents of Y (see structural equations 5.1) that are not X. Thus the causal estimand 5.8 is simply the total causal effect b_{XY}.

5.3.1. Regression for estimating linear causal effects

In general, the structural coefficient b_{XY} from the data generating (structural) model $Y = b_0 + b_{XY}X + \varepsilon$ need not correspond to the parameter β_{XY} estimated in a regression analysis of Y on X. To see that, note how the OLS estimator 3.5 approximates the quantity

$$\beta_{XY} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} = \frac{\mathrm{Cov}(X, b_{XY}X + \varepsilon)}{\mathrm{Var}(X)} = b_{XY} + \frac{\mathrm{Cov}(X,\varepsilon)}{\mathrm{Var}(X)} \tag{5.10}$$

by assuming $\mathrm{E}(\varepsilon \mid X) = 0$ (model 3.4 assumptions). If in fact $\mathrm{E}(\varepsilon \mid X) \neq 0$ then X and the error term ε are correlated and $\beta_{XY} \neq b_{XY}$. Since ε contains the effects of all variables other than X that act on Y, dependence $\mathrm{Cov}(X,\varepsilon) \neq 0$ can arise due to confounders of the X and Y relationship not accounted for in the regression model, but also due to reverse causation (Subsection 5.2.1).

For example, say we wanted to perform a regression of Y on X but the underlying data generating model was reverse causal: $X = b_{YX}Y + \varepsilon_X$, where $Y = \varepsilon_Y$ (assume centered variables for simplicity). The true causal effect of X on Y is $b_{XY} = 0$ but the OLS estimate of the regression analysis

$$\hat{\beta}_{XY} = \frac{\widehat{\mathrm{Cov}}(X,Y)}{\widehat{\mathrm{Var}}(X)} \cdot \frac{\widehat{\mathrm{Var}}(Y)}{\widehat{\mathrm{Var}}(Y)} = \hat{b}_{YX}\frac{\widehat{\mathrm{Var}}(Y)}{\widehat{\mathrm{Var}}(X)}$$

indicates $\hat{\beta}_{XY} \neq 0 \iff \hat{b}_{YX} \neq 0$. It means that ruling out reverse causation with OLS regression is impossible without domain knowledge. However, say the effect of Y on X went through a mediator M, such that $X = b_{M\to X}M + \varepsilon_X$, where $M = b_{Y\to M}Y + \varepsilon_M$. Accounting for M as a covariate in the regression model results in the following OLS estimator 3.5:

$$\hat{\beta}_{XY} = \frac{\widehat{\mathrm{Var}}(M)\overbrace{\widehat{\mathrm{Cov}}(X,Y)}^{\hat{b}_{YX}\widehat{\mathrm{Var}}(Y)} - \overbrace{\widehat{\mathrm{Cov}}(X,M)}^{\hat{b}_{M\to X}\widehat{\mathrm{Var}}(M)}\overbrace{\widehat{\mathrm{Cov}}(M,Y)}^{\hat{b}_{Y\to M}\widehat{\mathrm{Var}}(Y)}}{\widehat{\mathrm{Var}}(M)\widehat{\mathrm{Var}}(X) - \left(\widehat{\mathrm{Cov}}(X,M)\right)^2} \approx 0,$$

where the approximation holds because the total effect 5.9 of Y on X estimated as part of the first term of the numerator is $b_{YX} = b_{M\to X}b_{Y\to M}$.

The effect estimand in the regression analysis can become biased also due to failure to include confounding factors in the model. To illustrate with an example,

consider the DAG in Figure 4a, corresponding to the following structural equations (again assume centered variables for convenience):

$$\begin{aligned} U &= \varepsilon_U \\ X &= U \cdot b_{U \to X} + \varepsilon_X \\ M &= X \cdot b_{X \to M} + \varepsilon_M \\ Y &= U \cdot b_{U \to Y} + X \cdot b_{X \to Y} + M \cdot b_{M \to Y} + \varepsilon_Y \end{aligned}$$

The true causal effect 5.9 of X on Y is $b_{XY} = b_{X \to Y} + b_{X \to M} b_{M \to Y}$. However, a regression of Y on X results in

$$\mathrm{E}(Y \mid X) = \big(\underbrace{\mathrm{E}(U \mid X) b_{U \to Y} + b_{X \to Y} + b_{X \to M} b_{M \to Y}}_{\beta_{XY}}\big) X.$$

Omitting the confounder U from the regression analysis (i.e. by fitting a model $Y = \beta_{XY} X + \varepsilon$) induces a correlation between X and the errors in the regression model, leading to a biased estimand $\beta_{XY} \neq b_{XY}$. However, controlling for U gives

$$\mathrm{E}(Y \mid X, U) = b_{U \to Y} U + (\underbrace{b_{X \to Y} + b_{X \to M} b_{M \to Y}}_{\beta_{XY}}) X,$$

where the regression estimand β_{XY} is equal to the causal effect b_{XY}.

Note that both reverse causation and confounding represent back-door paths from the exposure to the outcome of interest. To estimate the causal effect 5.8, these back-door paths need to be closed (Subsection 5.2.1). OLS regression can be used to identify linear causal effects by controlling for a sufficient set of variables $\boldsymbol{V}_s$ satisfying the back-door criterion (Section 5.2) [40].

5.4. Method of instrumental variables for linear causal effects not directly identifiable

The back-door criterion introduced in Section 5.2 gives us sufficient means for identifying the causal effect. Alas, it cannot always be enforced. For example, if U in Figure 4a should be unobserved, then the causal effect of X on Y could not be identified by applying the back-door criterion to observational data. However, if the direct effect did not exist ($b_{X \to Y} = 0$), the causal effect b_{XY} could be estimated as the product of b_{XM} and b_{MY} (see equation 5.9), both of which are identifiable based on the back-door criterion. The procedure of estimating causal effects through causal paths is referred to as front-door and can be applied to all causal effects expressible as a combination of individually identified mediating effects [97]. Similarly to the back-door criterion however, it requires knowledge of the graph structure which we do not have. When we can assume linear causal effects (implies monotonicity), a more robust alternative to this assumption is provided by the method of instrumental variables.

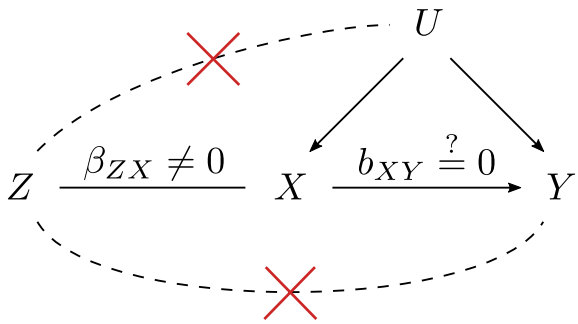


Figure 5: A graph depicting the assumptions of the method of instrumental variables for estimating the causal effect b_{XY} in the presence of unobserved confounders U of the exposure (X)-outcome (Y) relationship. Directed edges depict causal paths; undirected dashed edges represent any unblocked path; the solid undirected edge between Z and X represents an unblocked path that is not a causal path from X. In translation, a valid instrument Z needs to be associated with X (i.e. $\beta_{ZX} \neq 0$) and can be associated to Y only through a causal path from X.

Compared to the OLS regression 3.4 which enforces $E(\varepsilon \mid X) = 0$, the method of instrumental variables allows to identify the causal effect b_{XY} in the linear model $Y = b_0 + b_{XY}X + \varepsilon$ under a different (sometimes more plausible) set of assumptions. Indeed, consider an additional random variable Z that satisfies the following criteria [100, 101] (Figure 5):

(*relevance*) $\mathrm{Cov}(Z,X) \neq 0$, i.e. Z is associated with the exposure X; in graphical terms, there has to be an unblocked path between them

(*exogeneity*) $\mathrm{E}(\varepsilon \mid Z) = 0 \implies \mathrm{E}(\varepsilon) = 0 \wedge \mathrm{Cov}(Z,\varepsilon) = 0$, i.e. Z is unrelated to the error term ε; since the latter harbours all factors other than X that affect Y (incl. confounders U), this requirement is usually replaced by the following graphical assumptions [102–104]:

- (*exchangeability*) $Z \perp\!\!\!\perp U$, i.e. Z is independent (d-separated) from U
- (*exclusion restriction*) $Z \perp\!\!\!\perp Y \mid (X,U)$, i.e. there is no path between Z and Y that does not go through X

In short, Z needs to be associated with X and can be connected to Y only through a causal path from X. Any variable Z satisfying the above assumptions—we will refer to such variables as instruments—can be used to recover the causal effect b_{XY} between X and Y even in the presence of unobserved confounders between these variables [100, 101, 105]. Indeed, taking the covariance between the instrument Z and the outcome Y yields the following causal effect estimand 5.8 (remember that we still assume linear effects here; see [106] for a more general treatment of this estimand):

$$b_{XY} = \frac{\mathrm{Cov}(Y,Z)}{\mathrm{Cov}(X,Z)}. \tag{5.11}$$

It is straightforward to estimate the causal effect from a random sample by substi-

tuting true covariances with corresponding sample estimates as follows:

$$\hat{b}_{XY} = \frac{\widehat{\mathrm{Cov}}(Y,Z)}{\widehat{\mathrm{Cov}}(X,Z)} = \frac{\hat{\beta}_{ZY}}{\hat{\beta}_{ZX}} =: \hat{\beta}_{IV}, \tag{5.12}$$

where $\hat{\beta}_{IV}$ is called the IV estimator of causal effect between the exposure X and outcome Y [100, 101, 105] while $\hat{\beta}_{ZY}$ and $\hat{\beta}_{ZX}$ are OLS estimates 3.5 from the regressions of Y on Z and X on Z, respectively. Thus a valid instrument permits to estimate the causal effect as a ratio of two regression coefficients. Importantly, establishing causality boils down to testing for the instrument-outcome association, $H_0 : \beta_{ZY} = 0$. The IV analysis can thus be considered a computational extension to RCTs—often infeasible in practice as discussed in Section 1.2—as a valid instrument can be thought to represent an experimenter randomly allocating study participants into cases and controls.

5.4.1. The IV estimator is consistent and asymptotically normal

Before studying the properties of the IV estimator 5.12, we will assume without loss of generality that the instrument Z is centered (if not, we can always do that without changing the values of $\hat{\beta}_{ZY}$ or $\hat{\beta}_{ZX}$ as per Subsection 3.2.1) and has finite variance. Furthermore, we assume to have an independent and identically distributed random sample $(Z_i, X_i, Y_i) = (z_i, x_i, y_i)$, $i \in \mathcal{I}_n$, with homoscedastic errors $\mathrm{E}(\varepsilon_i^2 \mid Z_i) = \sigma_\varepsilon^2$.

First, note how the application of LIE leads to the following properties:

$$\begin{aligned} \mathrm{E}(Z_i\varepsilon_i) &= \mathrm{E}\big(Z_i\mathrm{E}(\varepsilon_i \mid Z_i)\big) = 0, \\ \mathrm{Var}(Z_i\varepsilon_i) &= \mathrm{E}\left(Z_i^2\mathrm{E}(\varepsilon_i^2 \mid Z_i)\right) = \sigma_\varepsilon^2\mathrm{Var}(Z). \end{aligned}$$

Making use of these properties and applying the law of large numbers (LLN), central limit theorem (CLT) and Slutsky's theorem enables us to show that the IV estimator 5.12 is consistent and asymptotically normally distributed:

$$\sqrt{n}(\hat{\beta}_{IV} - \beta_{IV}) = \Bigg(\underbrace{\frac{1}{n}\sum_{i=1}^{n} Z_iX_i}_{\downarrow \mathrm{Cov}(Z,X)}\Bigg)^{-1} \overbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i\varepsilon_i}^{\mathcal{N}\left(0,\,\sigma_\varepsilon^2\mathrm{Var}(Z)\right)} \xrightarrow{\;d\;} \mathcal{N}\Bigg(0, \underbrace{\frac{\sigma_\varepsilon^2\mathrm{Var}(Z)}{(\mathrm{Cov}(Z,X))^2}}_{\| \; \frac{\sigma_\varepsilon^2}{\rho_{ZX}^2\mathrm{Var}(X)}}\Bigg), \tag{5.13}$$

where ρ_{ZX}^2 denotes for correlation squared between Z and X. Furthermore, all terms of the asymptotic variance of $\hat{\beta}_{IV}$ are easily estimable: $\mathrm{Var}(X)$ as the sample variance of X, ρ_{ZX}^2 as the coefficient of determination R_{ZX}^2 from the regression of X on Z, and σ_ε^2 as the residual variance:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_{IV}x_i)^2.$$

Using these estimates, the hypothesis $H_0 : b_{XY} = 0$ can be tested with a Wald test.

5.4.2. Generalization of IV to multiple instruments

The IV estimator 5.12 is restrictive in the sense that it allows us to use only one instrument whereas sometimes we might have multiple instrumental variables at our disposal. It would be reasonable to ask whether we could improve our estimate of the causal effect by using all the information available to us.

Let $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_m)$ be a random vector of instruments and consider again the method of IV for estimating the causal effect b_{XY} from an exposure X to an outcome Y. Let $\beta_Z : m \times 1$ be a vector of weights such that $\boldsymbol{Z}'\beta_Z$ is a linear combination of the instrument vector. Since each Z_i, $i \in \mathcal{I}_m$ is an instrument, the linearity property of covariance ensures that $\mathrm{Cov}(\boldsymbol{Z}'\beta_Z, \varepsilon) = 0$. Thus $\boldsymbol{Z}'\beta_Z$ is also an instrument and can be used to estimate the causal effect with the IV estimator 5.12 whenever β_Z is such that $\mathrm{Cov}(\boldsymbol{Z}'\beta_Z, X) \neq 0$. The question remains which weight vector to use in constructing the linear combination.

As could be expected, a good strategy for instrument selection is to maximize its strength [101]; doing so simultaneously minimizes the (asymptotic) variance of the IV estimator (see relation 5.13 above):

$$\beta_Z^* = \underset{\beta_Z}{\mathrm{argmax}} \, |\mathrm{Cov}(\boldsymbol{Z}'\beta_Z, X)| = \underset{\beta_Z}{\mathrm{argmax}} \, \rho^2_{\boldsymbol{Z}'\beta_Z, X}, \tag{5.14}$$

where $\rho^2_{\boldsymbol{Z}'\beta_Z, X}$ denotes for correlation squared between $\boldsymbol{Z}'\beta_Z$ and X. Since each Z_i is an instrument, it follows trivially that the best linear combination satisfies $|\mathrm{Cov}(\boldsymbol{Z}'\beta_Z^*), X)| \geq |\mathrm{Cov}(Z_i, X)| \neq 0$ and is thus indeed itself an instrument.

The maximisation task 5.14 requires us to find a linear combination of $\boldsymbol{Z}$ that would explain the most variance in X. In a sample, this is equivalent to minimizing residual sum of squares in the regression of X on $\boldsymbol{Z}$. Estimates of optimal weights in 5.14 are thus conveniently provided by the OLS estimator 3.5:

$$\hat{\beta}_Z^* = \underset{\beta_Z}{\mathrm{argmax}} \, R^2_{\boldsymbol{Z}'\beta_Z, X} = (Z'Z)^{-1}Z'x, \tag{5.15}$$

where $R^2_{\boldsymbol{Z}'\beta_Z, X}$, $Z : n \times m$ and $x : n \times 1$ are sample equivalents of $\rho^2_{\boldsymbol{Z}'\beta_Z, X}$, $\boldsymbol{Z}$ and X, respectively. We can use the IV estimator 5.12 to estimate the causal effect:

$$\hat{\beta}_{IV} = \left((Z\hat{\beta}_Z^*)'x \right)^{-1} (Z\hat{\beta}_Z^*)'y = (\hat{x}'x)^{-1}\hat{x}'y, \tag{5.16}$$

where $\hat{x} = Z\hat{\beta}_Z^*$ are the fitted values of x from the OLS regression model. Since $Z(Z'Z)^{-1}Z'$ is idempotent, equation 5.16 further simplifies to

$$\begin{aligned} \hat{\beta}_{IV} &= \left(\hat{\beta}_Z^{*'} Z'Z\hat{\beta}_Z^* \right)^{-1} \hat{x}'y \\ &= (\hat{x}'\hat{x})^{-1}\hat{x}'y =: \hat{\beta}_{2SLS}. \end{aligned} \tag{5.17}$$

The resulting estimator $\hat{\beta}_{2SLS}$ has the form of an OLS estimator 3.5 and is called two stage least squares (2SLS) because it can be obtained in two regression steps [100]—the maximization task 5.15 is first solved for the prediction vector $\hat{x}$ and this is subsequently used as a regressor in 5.17 to estimate the causal effect.

5.4.3. Generalization of IV to multiple exposures

While the method of instrumental variables is promising for estimating causal effects, it requires the identification of valid instruments. Finding those may prove difficult however, particularly due to failure to satisfy the exogeneity assumption (unverifiable because model errors are unobserved). It might happen though that even if a potential instrument violates an assumption, it does so due to horizontal pleiotropy via other measured variables.

Consider a linear model $Y = \boldsymbol{X}' b_X + \varepsilon$, where $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$ is a random vector of exposures; and let $\boldsymbol{Z} : m \times 1$ be a random vector of instruments satisfying the following criteria: $\mathrm{Cov}(\boldsymbol{Z}, \boldsymbol{X})$ is of rank k and $\mathrm{E}(\varepsilon \mid \boldsymbol{Z}) = 0$ [101]. Analogously to the case with only one exposure, we can use the 2SLS procedure to estimate the causal effect vector b_X.

Let $\beta^*_{\boldsymbol{Z},X_i}$ be the vector of weights in 5.14 such that the linear combination $\boldsymbol{Z}'\beta^*_{\boldsymbol{Z},X_i}$ has the strongest association to the exposure X_i. Let $B^*_Z = \left(\beta^*_{\boldsymbol{Z},X_i} : i \in \mathcal{I}_k\right)$ be the corresponding $m \times k$ matrix with $\beta^*_{\boldsymbol{Z},X_i}$ in the columns. Estimating this matrix is straightforward as its every column is an OLS estimate 5.15, thus

$$\hat{B}^*_Z = (Z'Z)^{-1}Z'X,$$

where $Z : n \times m$ and $X : n \times k$ are sample (size n) equivalents of $\boldsymbol{Z}$ and $\boldsymbol{X}$, respectively. Following estimators 5.16 and 5.17,

$$\hat{\beta}_{IV} = (\hat{X}'X)^{-1}\hat{X}'y = (\hat{X}'\hat{X})^{-1}\hat{X}'y = \hat{\beta}_{2SLS}, \tag{5.18}$$

where $\hat{X} = Z\hat{B}^*_Z$ is an $n \times k$ matrix of fitted values of the exposure matrix X from the first stage regression. Since the IV estimator 5.18 is in the form of the OLS estimator 3.5, its variance is

$$\mathrm{Var}(\hat{\beta}_{2SLS}) = \sigma^2_\varepsilon (\hat{X}'\hat{X})^{-1}.$$

In estimating this variance, the error variance σ^2_ε can be approximated using a maximum likelihood estimator 3.8, though note that the residuals should be found using the observed values of the exposures (i.e. the design matrix $X : n \times k$) from the original linear model, not predicted values ($\hat{X} : n \times k$) from the first-stage regression (TWAS methods in Section 4.3 err against this). Furthermore, $\hat{X}'\hat{X}$ needs to be of full rank to be invertible for which $m \geq k$ is necessary—there needs to be at least as many instruments than exposures. The estimator $\hat{\beta}_{IV}$ remains consistent and asymptotically normal since LIE, LLN, CLT and Slutsky's theorem all generalize to random vectors [101].

The method of instrumental variables represents a simple yet elegant means to estimating the causal effect, boasting some nice statistical properties. It does not require knowledge of all structural relationships between variables like other causal inference methods exploiting the rules of d-separation (Subsection 5.2.1) [107–110]. The insensitivity to confounders of the exposure-outcome relationship also makes it robust to the key challenge facing ordinary linear regression. Consequently, it has become very popular in genetics and related fields where it is called Mendelian randomization (MR) due to using genetic variants—assumed to follow the principles of Mendelian inheritance—as instruments.

6. MENDELIAN RANDOMIZATION

> Truth ... is much too complicated to allow anything but approximations
>
> John von Neumann

The suitability of genetic variants as instruments in MR analyses hinges on the premise that genetic information passes down from parents to offspring following a systematic and inherently random process with no outside influences (Subsection 2.2.1). If true, genetic variants would be safe against confounding and reverse causation in association studies. However, population stratification and assortative mating can create differences in allele frequencies between groups of individuals (Subsection 2.2.1). If the trait distribution should differ between these groups then the association between genetic variants and respective traits can in fact be confounded by these very factors. This is important to keep in mind in MR analyses (spurious associations of genetic variants with traits could violate the assumptions for instruments in MR) but it is a general nuisance in genetic association studies which can be corrected for using statistical techniques [64, 111].

In effect, the usability of genetic variants as instruments is indeed promising. By extension, MR is theoretically sound and built on solid principles. It can even be thought of as nature's randomized trial [112]. However, genetic determinants of traits to be used as instruments are generally unknown to us and must be estimated using statistical techniques (Section 4.1). Furthermore, horizontal pleiotropy is extensive among complex traits and can invalidate the assumptions of instruments. While MR is not a magic solution to be used carelessly for these reasons, it has been utilized successfully to implicate reliable findings, such as the causal role of lipid traits on CVD risk [54].

6.1. Mendelian randomization estimator

Consider random variables G (genetic instrument), X (modifiable exposure such as gene expression), Y (outcome of interest), and U (confounder of the X-Y relationship). MR is a method of instrumental variables (see Figure 5 and substitute Z for G) where the genetic instrument G satisfies the IV assumptions of relevance and exogeneity (Section 5.4). Thus we can estimate b_{XY} as the ratio of regression estimates $\hat{\beta}_{GY}$ and $\hat{\beta}_{GX}$ from regressions of Y on G and X on G, respectively (see equation 5.12):

$$\hat{b}_{XY} = \frac{\hat{\beta}_{GY}}{\hat{\beta}_{GX}}. \tag{6.1}$$

It is clear from relation 5.13 that under MR assumptions $\hat{b}_{XY}$ is asymptotically consistent, converging in probability to b_{XY} as the sample size n grows [101]. However, due to the high cost of gene expression studies, available data with

complex trait measurements are scarce [88]. Therefore, individual level data in large sample sizes cannot be assumed. This necessitates the exploration of small sample properties of $\hat{b}_{XY}$. As it turns out, the MR estimator 6.1 is biased in finite samples [113] and exhibits low power [114].

6.1.1. Finite sample bias of the Mendelian randomization estimator

The estimator $\hat{b}_{XY}$ in equation 6.1 is a non-linear function $f : (X,Y) \mapsto X^{-1}Y$ of two random variables, thus it is not straightforward to derive its expected value analytically. However, we can use higher order methods to approximate [113]. A 2nd order Taylor expansion of the estimator around $\left(\mathrm{E}(\hat{\beta}_{GX}), \mathrm{E}(\hat{\beta}_{GY})\right)$ yields

$$\mathrm{E}(\hat{b}_{XY}) \approx \frac{\mathrm{E}(\hat{\beta}_{GY})}{\mathrm{E}(\hat{\beta}_{GX})} - \frac{\mathrm{Cov}(\hat{\beta}_{GY}, \hat{\beta}_{GX})}{\left(\mathrm{E}(\hat{\beta}_{GX})\right)^2} + \frac{\mathrm{Var}(\hat{\beta}_{GX})\mathrm{E}(\hat{\beta}_{GY})}{\left(\mathrm{E}(\hat{\beta}_{GX})\right)^3}.$$

To simplify the expression on the right-hand side, consider once more the assumptions of MR analysis. By the data generating (causal) model, the outcome is a linear function of the exposure: $Y = b_0 + b_{XY}X + \varepsilon_{XY}$. The exposure is not assumed to be exogenous in the model but the instrument is: $\mathrm{Cov}(G, \varepsilon_{XY}) = 0$. The instrument is also assumed to be correlated with the exposure and the linear relationship implied by this notion can be depicted by an OLS model 3.4 with the exogeneity assumption enforced: $X = \beta_{0X} + \beta_{GX}G + \varepsilon_{GX}$ with $\mathrm{Cov}(G, \varepsilon_{GX}) = 0$, where $\beta_{GX} \neq 0$ but not necessarily due to causality (e.g. confounding can be a factor, hence the difference in notation compared to the causal model). The outcome can be expressed in terms of the instrument as $Y = \beta_{0Y} + \beta_{GY}G + \varepsilon_{GY}$, where $\beta_{GY} = b_{XY}\beta_{GX}$. This is an OLS model 3.4 as the exogeneity assumption is satisfied: $\mathrm{Cov}(G, \varepsilon_{GY}) = \mathrm{Cov}(G, \varepsilon_{XY} + b_{XY}\varepsilon_{GX}) = 0$. Finally, treating the instrument as fixed and centered, and considering that OLS estimators are unbiased (equation 3.6) with variance 3.7, we can approximate the final sample bias of the MR estimator 6.1 as follows:

$$\begin{aligned}
\mathrm{E}(\hat{b}_{XY}) - b_{XY} &\approx -\frac{\mathrm{Cov}(\hat{\beta}_{GY}, \hat{\beta}_{GX})}{\beta_{GX}^2} + \frac{\mathrm{Var}(\hat{\beta}_{GX})b_{XY}}{\beta_{GX}^2} \\
&= -\frac{\mathrm{Cov}(\varepsilon_{GY}, \varepsilon_{GX})}{\beta_{GX}^2 n\widehat{\mathrm{Var}}(G)} + \frac{\sigma_{\varepsilon_{GX}}^2 b_{XY}}{\beta_{GX}^2 n\widehat{\mathrm{Var}}(G)} \\
&= -\frac{\mathrm{Cov}(\varepsilon_{XY}, \varepsilon_{GX}) + \sigma_{\varepsilon_{GX}}^2 b_{XY} - \sigma_{\varepsilon_{GX}}^2 b_{XY}}{\beta_{GX}^2 n\widehat{\mathrm{Var}}(G)} \\
&\approx -\frac{\mathrm{Cov}(\varepsilon_{GX}, \varepsilon_{XY})}{n\rho_{G,X}^2 \mathrm{Var}(X)},
\end{aligned} \tag{6.2}$$

where n is the sample size and $\rho_{G,X}^2$ is squared correlation between the instrument G and exposure X. Of course, the finite sample bias in the MR effect estimate is

undesirable. It is worth a closer inspection, when and in which circumstances the bias manifests or is substantial.

First, note that the covariance $\mathrm{Cov}(\varepsilon_{GX}, \varepsilon_{XY})$ in the numerator of the bias 6.2 depends on the level of confounding between X and Y, being zero only if there is no confounding. Thus whenever MR is actually needed, its estimate of causal effect is never unbiased in a finite sample [105]. We can think of this bias as a violation of the MR assumption that instruments G are not associated with confounders U of the X-Y relationship—even if true, chance correlations in finite samples can bias the MR causal effect estimate [103]. Theoretically, this bias can be even greater than the bias of the OLS effect estimate in equation 5.10. However, this is increasingly unlikely as the sample size n grows since even in the presence of counfounders, the bias of the MR effect estimate approaches zero while the bias of the OLS effect estimate does not.

Second, the bias 6.2 is inversely proportional to sample size n and instrument strength $\rho^2_{G,X}$. A weak instrument in a small sample can thus inflate the bias. This is even more evident if we consider that the F-statistic from the regression of X on G can be approximated as follows:

$$F_{G,X} = (n-1)\frac{R^2_{G,X}}{1-R^2_{G,X}} \approx nR^2_{G,X}, \tag{6.3}$$

since for a weak instrument the denominator is approximately equal to 1. Note that the approximate F-statistic 6.3 can be used to estimate $n\rho^2_{G,X}$ in the denominator of the bias 6.2. Thus increasing the F-statistic value $F_{G,X}$ for the association between G and X decreases the bias. Rule of thumb is to have $F_{G,X} > 10$, though it would be best if instruments are chosen from an independent sample [103, 115].

Finally, note that we derived the bias 6.2 only for a single instrumental variable. With many instruments the bias can be exacerbated [116] even though the asymptotic variance of the estimator would decrease as evident in relation 5.13. This can happen due to having to estimate the optimal linear combination instrument in the first stage regression 5.15 of the 2SLS procedure.

6.1.2. Statistical power of Mendelian randomization

Given that individual gene expression studies with complex trait measurements are usually available in small samples [88], it is prudent to estimate the sample size required for MR to achieve sufficient power for testing the null hypothesis $H_0 : b_{XY} = 0$. Power of a statistical test is just the probability to reject the null hypothesis—observe as extreme or more extreme test statistic values than $q_{\frac{\alpha}{2}}$ and $q_{1-\frac{\alpha}{2}}$ specified by the significance level α—conditional on the true causal effect being $b_{XY} \neq 0$. Considering the asymptotic results in 5.13, we have

$$P\left(\frac{\hat{b}_{XY}}{\sqrt{\mathrm{Var}(\hat{b}_{XY})}} \geq q_{1-\frac{\alpha}{2}} \mid b_{XY}\right) + P\left(\frac{\hat{b}_{XY}}{\sqrt{\mathrm{Var}(\hat{b}_{XY})}} \leq q_{\frac{\alpha}{2}} \mid b_{XY}\right) = \text{Power},$$

where $q_{\frac{\alpha}{2}}$ and $q_{1-\frac{\alpha}{2}}$ are quantiles of the standard normal distribution. In particular, we are interested in the smallest sample size n^* required to reject the null hypothesis on significance level α with probability at least Power. Being conservative and neglecting the finite sample bias of MR, we can derive n^* as

$$n^* = \min n \left| 1 - \Phi\left(q_{1-\frac{\alpha}{2}} - \frac{|b_{XY}|}{\sqrt{\text{Var}(\hat{b}_{XY})}} \right) + \Phi\left(q_{\frac{\alpha}{2}} - \frac{|b_{XY}|}{\sqrt{\text{Var}(\hat{b}_{XY})}} \right) \geq \text{Power}, \right.$$

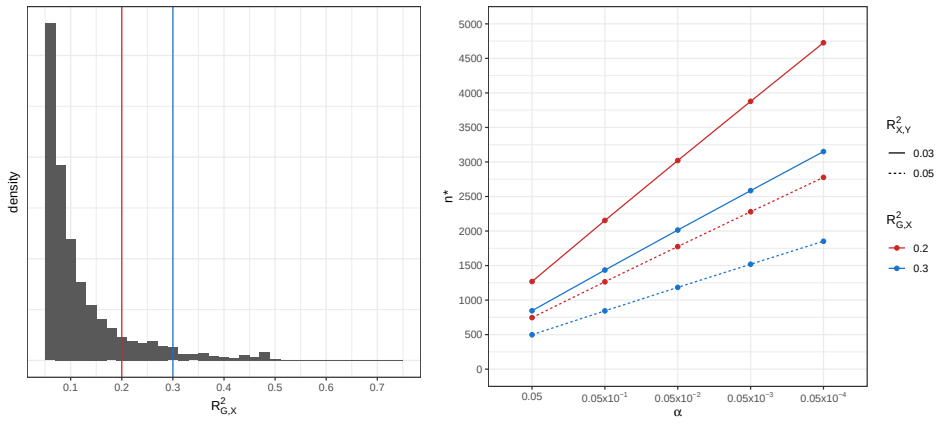
where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Thus n^* depends on the significance level α, specified Power, true effect size b_{XY}, and variance of the estimator $\hat{\beta}_{XY}$ in 5.13, the latter in turn depending on instrument strength $\rho^2_{G,X}$.

To estimate n^* in realistic scenarios, we need to come up with reasonable values for instrument strength $\rho^2_{G,X}$ and true causal effect b_{XY}; usually $\alpha = 0.05$ for a single test and Power $= 0.8$. Based on whole blood *cis*-eQTLs of the eQTLGen Consortium data [69] (https://www.eqtlgen.org/), the mean strength of potential gene expression instruments is $\bar{R}^2_{G,X} = 0.013$ (Figure 6a). However, since multiple instruments can be combined to increase the strength (see 5.14) and we prefer to be conservative in our calculations, we will consider very strong instruments describing 20% and 30% of the variability of gene expression traits, and strong causal effects with exposures describing 3% and 5% of outcome variability. Evidently, the required sample size n^* even for a very strong instrument and causal effect is approximately 500 for a single test (Figure 6b). Considering there are tens of thousands of genes and thus a necessity for multiple testing correction in hypothesis-free analysis, we would need thousands of samples to identify causal effects even in the best-case scenarios [114].

Considering low statistical power (Figure 6) and the small sample bias 6.2, MR analysis is feasible only in large samples. It would almost certainly be unfruitful in individual level data concerning gene expression measurements (for reference, the Estonian Biobank with more than 200,000 participants has gene expression RNA sequencing (RNA-seq) data on approximately 500 individuals [1]). Fortunately, we will see that MR can be performed on summary statistics meta-analyzed together over many individual cohorts with exposures and outcomes measured in independent samples.

6.2. Two-sample Mendelian randomization

Remember that the MR estimator 6.1 is a ratio of two regression coefficients. We know from Subsection 3.1.2 that for both of those we can use published summary statistics which are the result of meta-analysis of several individual estimates (using equation 3.11) conducted by large international consortia. Furthermore, $\hat{\beta}_{GY}$ and $\hat{\beta}_{GX}$ can be calculated from different studies, leading to *two-sample*



(a) (b)

Figure 6: (a) Distribution of instrument strength $R^2_{G,X}$ based on significant *cis*-eQTLs in the eQTLGen Consortium whole blood gene expression data [69] (https://www.eqtlgen.org/), filtered by $R^2_{G,X} \geq 0.05$ to improve readability. Instrument strength was calculated as $R^2_{G,X} = \left(\hat{\beta}^s\right)^2$ where $\hat{\beta}^s$ was estimated based on equation 3.14 using Z-scores and sample sizes from the eQTLGen data. (b) Minimum sample size n^* required per significance level α (corresponding to 1, 10, 10^2, 10^3, and 10^4 independent tests) to achieve 80% power in different instrument strength $R^2_{G,X}$ and causal effect $R^2_{X,Y}$ pairs. Sample size calculations use asymptotic consistency and normality of the MR estimator (see 5.13) and are therefore conservative approximations.

MR [117]. There are obvious benefits to this strategy. First, meta-analyzed summary statistics leverage information from several cohorts, achieving estimation precision proportional to the sum of sample sizes of individual studies (consider that each individual variance 3.7 in equation 3.12 is approximately equal to the inverse of the sample size if genetic effects are tiny and we assume standardized variables). Second, if each of the regression coefficients $\hat{\beta}_{GY}$ and $\hat{\beta}_{GX}$ have been calculated on independent samples then $\mathrm{Cov}(\varepsilon_{GX}, \varepsilon_{XY}) = 0$ in the 2nd order MR bias approximation 6.2, decreasing the overall bias [116]. This independence can indeed be assumed in transcriptomics analyses since the biggest gene expression studies made available in the public domain by the Genotype-Tissue Expression project (GTEx) [118] and the eQTLGen Consortium [69] have been conducted independently from large-scale analyses of complex traits, such as those based on the UK Biobank (UKBB) resource [119, 120]. Thus two-sample MR based on summary statistics alleviates the problems of low power and finite sample bias of MR and is therefore widely used in statistical genetics.

In this work, we are predominantly interested in the causal effect of gene expression X on complex trait Y. For this reason—without loss of generality—we refer to instruments as eQTLs. Let $\hat{\beta}_{G_iX}$ and $\hat{\beta}_{G_iY}$ be summary statistics from the

eQTL and GWAS analyses with genetic variants G_i, $i \in \mathcal{I}_m$, respectively. By the MR estimator 6.1, the causal effect estimate for each G_i is

$$\hat{\beta}_{IV,i} = \frac{\hat{\beta}_{G_iY}}{\hat{\beta}_{G_iX}}.$$

If G_i are indeed valid instruments then the application of LLN and Slutsky's theorem in 5.13 reveals that each estimate $\hat{\beta}_{IV,i}$ is consistent. The variance of the two-sample MR estimator can be approximated with the Delta method as follows (see proof in Appendix A):

$$\begin{aligned}\mathrm{Var}\left(\hat{\beta}_{IV,i}\right) &\approx \frac{\mathrm{Var}(\hat{\beta}_{G_iY})}{\beta^2_{G_iX}} - \frac{2\beta_{G_iY}\mathrm{Cov}(\hat{\beta}_{G_iY}, \hat{\beta}_{G_iX})}{\beta^3_{G_iX}} + \frac{\beta^2_{G_iY}\mathrm{Var}(\hat{\beta}_{G_iX})}{\beta^4_{G_iX}} \\ &= \frac{\mathrm{Var}(\hat{\beta}_{G_iY})}{\beta^2_{G_iX}} + \frac{\beta^2_{G_iY}\mathrm{Var}(\hat{\beta}_{G_iX})}{\beta^4_{G_iX}}, \end{aligned} \tag{6.4}$$

where $\mathrm{Cov}(\hat{\beta}_{G_iY}, \hat{\beta}_{G_iX}) = 0$ follows from the fact that $\hat{\beta}_{G_iY}$ and $\hat{\beta}_{G_iX}$ were assumed to be estimated in independent samples. We can estimate the variance 6.4 as

$$\widehat{\mathrm{Var}}(\hat{\beta}_{IV,i}) \approx \frac{\widehat{\mathrm{Var}}(\hat{\beta}_{G_iY}) + \left(\hat{\beta}_{G_iY}/\hat{Z}_{G_iX}\right)^2}{\hat{\beta}^2_{G_iX}}, \tag{6.5}$$

where $\hat{Z}_{G_i,X}$ is the Z-score from the regression of X on G_i. We have thus all the necessary machinery for estimating causal effects in the MR framework. However, it turns out identifying valid instruments is not as straightforward.

6.2.1. A simple fine-mapping strategy for a single instrument MR

Any SNP satisfying the relevance condition (achieving a test statistic value over some threshold in an eQTL study) could be a potential instrument. Due to LD there are likely to be many SNPs to choose from and we will not know the true causal variant(s). This is not in and of itself a problem since the relevance condition of MR does not actually state that instruments should be causally related to the exposure. Theoretically we could even use all the (correlated) eQTLs as instruments provided they satisfied MR assumptions, though this could lead to increased finite sample bias and would not help with statistical power if there was actually only one causal variant [121]. However, LD patterns can differ between samples and this poses a problem for two-sample analyses—a non-causal SNP satisfying the relevance condition in the eQTL study due to LD with the true causal SNP can exhibit either less or more LD in the GWAS, leading to biased causal effect estimates and spurious conclusions. Identifying true causal variants is therefore desirable.

In line with fine-mapping strategies in Section 4.1, the simplest MR approach is to assume one causal variant and let it be the strongest one with the lowest eQTL

P-value, a method referred to as summary data-based Mendelian randomization (SMR) [88]:

$$\hat{\beta}_{SMR} = \frac{\hat{\beta}_{G_*Y}}{\hat{\beta}_{G_*X}}, \quad G_* = \underset{G_i : i \in \mathcal{I}_m}{\operatorname{argmax}} |\hat{Z}_{G_i,X}|. \tag{6.6}$$

Compared to other potential instruments, choosing the strongest eQTL G_* is likely to result in the smallest variance of the two-sample MR estimator in equation 6.5. The variance is not definitively minimized only because it also depends on GWAS summary statistics and hence sample sizes which can vary between SNPs. Nevertheless, G_* is a good candidate for increasing the certainty of the causal effect estimate. However, one should be aware that the effect size of the strongest eQTL is likely to be overestimated due to random chance—a phenomenon called the winner's curse [122]—which in a two-sample analysis biases the causal effect estimate $\hat{\beta}_{SMR}$ towards the null (due to $\hat{\beta}_{G_*X}$ being in the denominator) [123].

6.2.2. Allowing for multiple instruments

Though the SMR approach is enticing in its simplicity, we know that there can be many causal variants responsible for trait variability and using them all could further improve the precision of the causal effect estimate. To prioritize multiple causal variants, we can use other fine-mapping approaches covered in Section 4.1, such as stepwise conditional analysis.

Let G_i, $i \in \mathcal{I}_m$ be all the identified instruments with corresponding causal effect estimates $\hat{\beta}_{IV,i}$. To combine these into a single estimate, it is straightforward to use the meta-analysis theory developed in Subsection 3.1.2. If G_i can be assumed to not be in LD—and we have already covered that it makes sense to select independent instruments—then we can apply weighted linear regression on model 3.10 to estimate the causal effect with the IVW average 3.11 [117]:

$$\hat{\beta}_{IVW} = \frac{\sum_{i=1}^{m} \left(\mathrm{Var}(\hat{\beta}_{IV,i})\right)^{-1} \hat{\beta}_{IV,i}}{\sum_{i=1}^{m} \left(\mathrm{Var}(\hat{\beta}_{IV,i})\right)^{-1}}, \tag{6.7}$$

where $\mathrm{Var}(\hat{\beta}_{IV,i})$ can be substituted by approximations 6.4 (however, assuming $\mathrm{Var}(\hat{\beta}_{G_iX}) = 0$ can result in better properties like smaller bias [124]).

For uncorrelated instruments, the IVW estimator 6.7 is asymptotically equivalent to the 2SLS estimator 5.17 [125]. Sometimes LD between instruments cannot be ruled out however, particularly in gene expression studies where several *cis*-acting eQTLs can be close to each other. Discarding correlated instruments would be an option in this case but not an optimal one. Fortunately, model 3.10 can be generalized to account for LD by allowing for correlations between effect estimates (these can be found with the Delta method, see Appendix A). Such method is termed Generalized SMR [126]. Similarly to weighted least squares, the generalized model can be solved by reducing it to OLS regression.

6.2.3. TWAS-like polygenic score instruments for causal inference

Instead of combining causal effect estimates into a weighted average, we could also aggregate individual instruments before applying the simple MR estimator 6.1. Indeed, we know that a linear combination of instruments is also an instrument (Subsection 5.4.2). A PRS of the exposure can thus be used to estimate its causal effect on the outcome. In fact, the individual level data-based TWAS estimator 4.13 is exactly identical to the IV (and thus MR) estimator 5.16. Therefore, the summary level data-based TWAS estimator 4.15 with corresponding variance 4.16 can also be used for MR analysis.

As we have shown, TWAS and MR are fundamentally the same. This has been corroborated by other studies [127]. However, recall that we paradoxically claimed TWAS to be an *ad hoc* method for causal inference, invalid even (Subsection 4.3.1). Like other PRS-based association approaches, TWAS simply does not approach the problem of causal inference with the required level of theoretical soundness and due diligence. For example, TWAS methods [50–52] are not concerned with instrument assumptions (Section 5.4), the optimal way of combining instruments (Subsection 5.4.2), nor bias from weak instruments (Subsection 6.1.1). Simply put, invalid instruments combined to a PRS represent an invalid instrument [128] and do not enable causal reasoning.

To interpret TWAS results in a causal language, all the theory developed for MR applies and should be adhered to. In this regard, MR subsumes TWAS and similar PRS association-based approaches alike. The applicability of causal reasoning reduces to avoiding horizontal pleiotropy in the instruments.

6.3. Pleiotropy in Mendelian randomization

To estimate the causal effect b_{XY} of X on Y in a linear model $Y = b_0 + b_{XY}X + \varepsilon$ with the MR estimator 6.1, the instrument G needs to satisfy the conditions of relevance and exogeneity: $\text{Cov}(G,X) \neq 0$ and $\text{Cov}(G,\varepsilon) = 0$, respectively (Section 5.4). As seen in the previous two subsections, identifying instruments in terms of satisfying the relevance condition can be performed in a data-driven manner using fine-mapping strategies. However, since errors ε are unobserved, it is not as straightforward to verify that selected instruments are valid also in terms of satisfying the exogeneity assumption; or equivalently do not affect Y through paths not consisting X (see Figure 5). Violation of this requirement suggests pleiotropy. Since pleiotropy is known to be pervasive [41], it represent a serious hazard for the reliability of MR.

It is clear from the convergence 5.13 that failure to satisfy $\text{Cov}(G,\varepsilon) = 0$ introduces a bias to the causal effect estimator 6.1. This bias cannot be partitioned from the true causal effect. In accordance with the theory of causal inference, a significant effect $\hat{b}_{XY}$ in an MR analysis for any single SNP G (e.g. consider the SMR method 6.6) can thus arise due to the following reasons or combinations

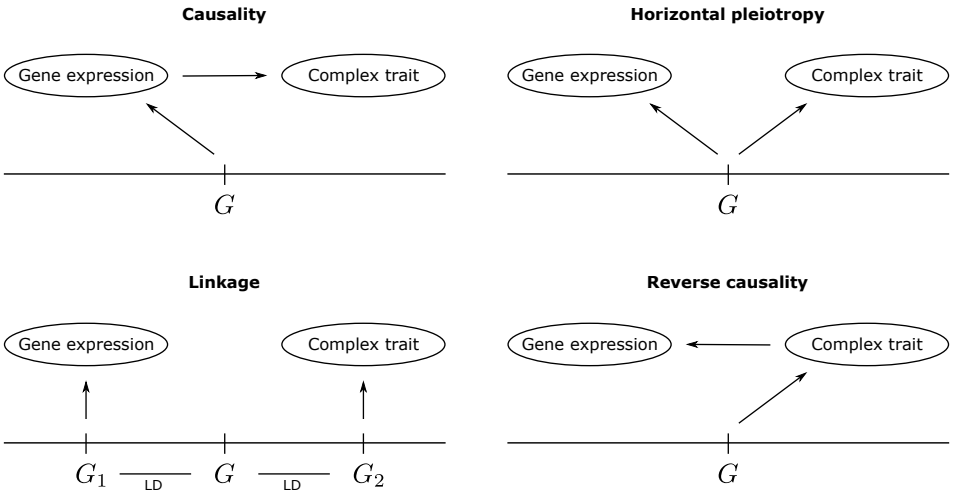


Figure 7: Possible explanations for observing a significant effect in a Mendelian randomization analysis with a single instrument G.

thereof (Figure 7):

(i) Causality—there is a causal path from the exposure to the outcome.

(ii) Horizontal pleiotropy—G affects the outcome through paths not mediated by the exposure.

(iii) Linkage—two different SNPs, both in LD with G, are responsible for the effects on the exposure and outcome.

(iv) Reverse causality—the effect on the exposure is mediated by the outcome instead, though this is unlikely in gene expression studies where eQTLs are used as instruments.

Thus if all we have is one instrument, determining causality with standard MR is impossible without prior knowledge. Note however that using *cis*-eQTL instruments makes reverse causality unlikely. Furthermore, discriminating linkage from causality and horizontal pleiotropy—both of which are based on the assumption of a single causal variant—is possible with colocalization methods introduced in Section 4.2. The HEIDI test 4.11 in particular has been developed for this purpose and can be used to test for the null hypothesis of a shared causal variant [88]. Hence, combining SMR with colocalization methods provides a ranking of likely causal genes. This can be sufficient if the goal is simply to prioritize genes for further analyses. If determination of the true causal effect size is required, more instruments are needed [129].

6.3.1. Determining pleiotropy in multi-instrument setting

Consider m genetic variants G_i, $i \in \mathcal{I}_m$. Let β_{G_iX} be the effect of G_i on the exposure X and let ω_i be its effect on the outcome Y through paths not consisting X. Both the exposure and outcome can be written in terms of each instrument using the

following OLS models (see Subsection 6.1.1 for reference):

$$
\begin{aligned}
X &= \beta_{0X} + \beta_{G_iX} G_i + \varepsilon_{G_iX}, \\
Y &= b_0 + b_{XY} X + \omega_i G_i + \varepsilon_{XY} = \beta_{0Y} + \underbrace{(b_{XY}\beta_{G_iX} + \omega_i)}_{\beta_{G_iY}} G_i + \varepsilon_{G_iY},
\end{aligned}
$$

where b_{XY} is the causal effect of X on Y and β_{G_iY} is the effect (not necessarily causal) of G_i on Y. In MR analysis, we are estimating the following quantity with each instrument (see equation 5.11):

$$
\beta_{IV,i} = \frac{\mathrm{Cov}(Y, G_i)}{\mathrm{Cov}(X, G_i)} = \frac{\beta_{G_iY}}{\beta_{G_iX}} = b_{XY} + \frac{\omega_i}{\beta_{G_iX}},
$$

which is biased for the true causal effect if $\omega_i \neq 0$. This is exactly the case with pleiotropic instruments. Unfortunately, elucidating the bias caused by non-zero ω_i is not possible without knowing the underlying structural model of the variables involved. Since pleiotropy is known to be widespread [41, 130], simply assuming $\omega_i = 0$ is often not justified. To work around this issue, we can turn to the theory of meta-analysis introduced in Subsection 3.1.2.

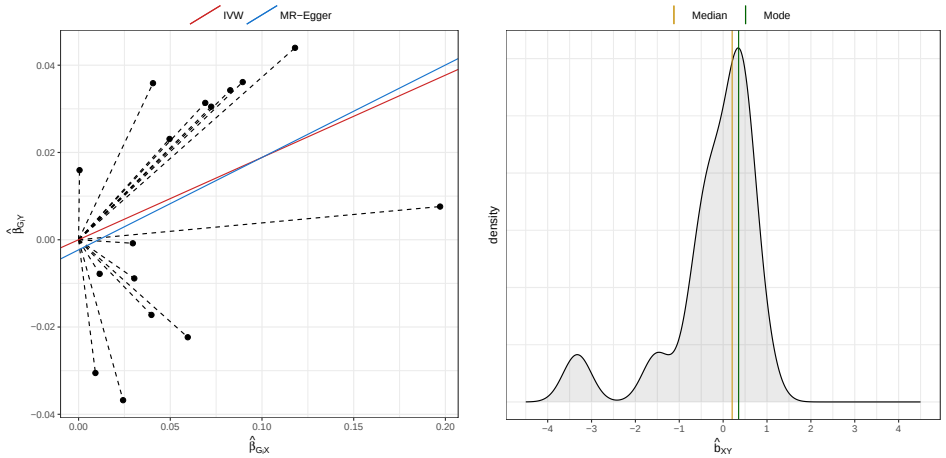
Let instruments G_i be independent and consider the IVW average 6.7 for estimating the causal effect. In the absence of pleiotropy, the true effects $\beta_{IV,i}$ are identical for all $i \in \mathcal{I}_m$. The homogeneity in effects means that we could construct $\hat{\beta}_{IVW}$ using weights that are simple inverses of study-specific variances 6.4. Validity of the homogeneity assumption can be tested with the Cochran's Q statistic 3.13 and, in case of heterogeneity, we could either account for study-specific biases with the DerSimonian and Laird method, or eliminate pleiotropic instruments from the analysis altogether (Subsection 3.1.2). Several recent methods are based on removing outlying instruments, such as MR pleiotropy residual sum and outlier (MR-PRESSO) test [130] and HEIDI-outlier test [126]. The latter determines outliers on the basis of the $\hat{d}_i$-statistic 4.10 being significantly different from its expected value. It is prudent to note that the sum of $\hat{d}_i$, the T_{HEIDI} statistic 4.11—developed for detecting effect heterogeneity in colocalization analyses (Subsection 4.2.2)—can also be used to detect pleiotropy-induced heterogeneity in MR studies, thus generalizing the Cochran's Q statistic for dependent instruments.

Under the assumption that biases in causal effect estimates of individual instruments cancel out, $\mathrm{E}\left(\frac{\omega_i}{\beta_{G_iX}}\right) = 0$, the corresponding IVW estimate $\hat{\beta}_{IVW}$ is consistent for the true causal effect even if all the instruments are pleiotropic. If instrument strengths β_{G_iX} are independent from direct/pleiotropic effects ω_i (abbreviated as the InSIDE assumption), we could equivalently assume zero average pleiotropy, since in this case $\mathrm{E}\left(\frac{\omega_i}{\beta_{G_iX}}\right) = 0 \iff \mathrm{E}(\omega_i) = 0$ [104]. It is therefore not strictly necessary in multi-instrument two-sample MR settings to require all instruments to be valid—it is sufficient for obtaining consistent causal effect estimates if pleiotropic effects cancel out under the InSIDE assumption.

6.3.2. Additional sensitivity analyses

Recall that the IVW estimator $\hat{\beta}_{IVW}$ is the solution of an intercept-free weighted least squares regression of individual causal effect estimates (Subsection 3.1.2) and thus represents a slope of the regression line which goes through the origin (Figure 8a). In the presence of directional pleiotropy, if the average pleiotropic effect of instruments do not cancel out, the slope will be biased. To account for directional pleiotropy, we can turn to Egger regression (introduced in Subsection 3.1.2) and allow for the intercept in model 3.10, a technique which in the context of MR is termed MR-Egger regression [102]. Similarly to $\hat{\beta}_{IVW}$, the causal effect estimate from MR-Egger regression is consistent even when all instruments are invalid, provided the InSIDE assumption holds [102]. However, even though MR-Egger can correct for directional pleiotropy, there is little power to do so if the number of instruments is small or the instruments are collectively weak [131–133].

Instead of estimating the causal effect with IVW or MR-Egger regressions—through a linear combination of individual estimates $\hat{\beta}_{IV,i}$—complimentary measures of central tendency can be considered for the same purpose. For example, median [132] or mode [134] of the empirical distribution of $\hat{\beta}_{IV,i}$ are good candidates for representing the causal effect (Figure 8b). These would provide consis-



(a) (b)

Figure 8: Sensitivity analyses to detect horizontal pleiotropy in a Mendelian randomization study. (a) IVW and MR-Egger causal effect estimates (red and blue slopes, respectively) together with individual estimates from independent instruments (slopes of dashed black lines), based on data in **Ref. I** [1]. The similarity of IVW and MR-Egger estimates indicates no directional pleiotropy among the instruments. (b) Distribution of causal effect estimates of individual instruments with corresponding median (yellow) and mode (green).

tent estimates if at least 50% of the instruments were valid or if the most frequent was valid, respectively. Since the validity of any of the assumptions governing pleiotropy cannot be verified, using orthogonal methods in the form of sensitivity analyses is likely to provide more reliable causal effect estimates compared to using just a single method [129, 132].

6.4. Multivariable Mendelian randomization

Instead of attempting to correct for horizontal pleiotropy by relying on untestable assumptions or discarding invalid instruments, we could try to account for all possible causal paths from genetic variants to the outcome by including potential risk factors in the model simultaneously (Figure 9). MR can easily be generalized like that (Subsection 5.4.3). Since the widespread physiological pleiotropy seen among complex traits is likely to originate from pleiotropy in the regulatory level [135], a multivariable Mendelian randomization (MVMR) approach looks especially promising in gene expression studies.

Let Y be a random variable of a complex trait outcome, $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ a random vector of gene expression exposures and $\mathbf{G} = (G_1, G_2, \ldots, G_m)$ a random vector of genetic instruments. For simplicity, assume all the variables are centered and standardized. Now consider the MVMR model $Y = \mathbf{X}b + \varepsilon$ with $\mathrm{E}(\varepsilon \mid \mathbf{G}) = 0$ (Subsection 5.4.3). The MVMR causal effect can be estimated by the two-stage least squares estimator 5.18 by incorporating all of the instruments and exposures in the model at the same time. However, this approach requires individual-level data and thus is often not applicable in genetic studies. Recently, McDaid et al. derived a summary statistics-based solution for estimating the MVMR causal effect [136]:

$$\hat{b} = \left(\hat{\Gamma}' C^{-1} \hat{\Gamma}\right)^{-1} \hat{\Gamma}' C^{-1} \hat{\gamma}, \tag{6.8}$$

where $\hat{\Gamma} : m \times k$ and $\hat{\gamma} : m \times 1$ are standardized OLS effect estimates of genetic instruments on exposures and outcome, respectively; and $C : m \times m$ is an LD-correlation matrix of genetic instruments which can be estimated using reference data. I have derived a covariance matrix for the MVMR causal effect estimator 6.8 using the Delta method [3] (see Appendix A for all the derivation details):

$$\mathrm{Var}(\hat{b}) = J(\beta)\mathrm{Var}(\hat{\beta})J(\beta)',$$

where $\hat{\beta} = \left(\mathrm{vec}(\hat{\Gamma}) \mid \hat{\gamma}'\right)'$ is the vector of individual effect estimates with true effects $\beta = \mathrm{E}(\hat{\beta})$, $\mathrm{Var}(\hat{\beta})$ is the corresponding variance-covariance matrix which can be estimated using summary statistics and a genotype reference, and

$$J(\beta) = \left(\frac{\partial \hat{b}}{\partial \hat{\Gamma}}(\beta) \;\middle|\; \frac{\partial \hat{b}}{\partial \hat{\gamma}}(\beta) \right)$$

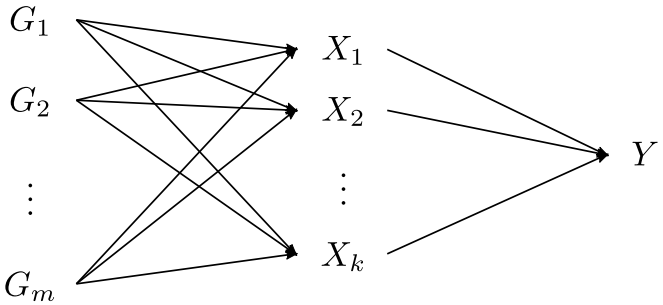


Figure 9: Schematic of multivariable Mendelian randomization with an outcome Y, k exposures ($X_1, X_2, \ldots, X_k$), and m genetic instruments ($G_1, G_2, \ldots, G_m$).

is the Jacobian matrix of $\hat{b}$ evaluated at β (our notation is slightly loose here but note that $\hat{b}$ is essentially a function $f : \mathbb{R}^{m(k+1)} \to \mathbb{R}^k$ and thus takes any $m(k+1)$-vector as an argument). However, the true effects are unknown, thus we will approximate the Jacobian at $\hat{\beta}$ instead. Since the estimates are already naturally incorporated in $\hat{\Gamma}$ and $\hat{\gamma}$, we can omit the argument altogether in the derivatives:

$$\begin{aligned}
\frac{\partial \hat{b}}{\partial \hat{\Gamma}} &= \left((\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \otimes \left(- \hat{\gamma}' C^{-1} \hat{\Gamma} (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \hat{\Gamma}' C^{-1} \right)\right) + \\
&\quad + \left(- \hat{\gamma}' C^{-1} \hat{\Gamma} (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \otimes (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \hat{\Gamma}' C^{-1} \right) + \\
&\quad + \left((\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \otimes (\hat{\gamma}' C^{-1}) \right), \\
\frac{\partial \hat{b}}{\partial \hat{\gamma}} &= (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \hat{\Gamma}' C^{-1},
\end{aligned}$$

giving us everything we need to estimate the variance-covariance matrix of the causal effect vector.

The difficulty in MVMR arises from having to identify all possible exposures (genes) through which the effect of any given instrument on the outcome could propagate. MVMR controls for pleiotropic effects through these exposures under investigation [54], estimating exposure effects on the outcome that are not mediated by other risk factors in the model [137]. This is different from single-exposure MR which estimates the total causal effect of the risk factor on the outcome, including the effect mediated by genes in the MVMR model. Thus causal effects from univariable and multivariable MR do not have to be in agreement even if there is no violation of MR assumptions. This could be an important consideration for some practical applications. Our quest to identify drug targets means we are interested in the total causal effect of genes. However, the dynamics between genes are hard to disentangle and single-gene MR analyses are likely to suffer substantially more due to pleiotropy—a bigger challenge to overcome than adjusting causal effects of prioritized genes later on in follow-up analyses. MVMR is thus well suited for transcriptomics studies (**Ref. III**), though can benefit from being complemented with single-gene MR (**Ref. II**) and sensitivity analyses.

6.4.1. Dealing with remaining heterogeneity in effect estimates

MVMR can effectively deal with pleiotropic effects propagating through exposures selected in the model. However, causal effect estimates can still be biased due to failure to control for all the assumptions (Subsection 5.4.3); unmeasured confounders in particular remain problematic. Additional sensitivity analyses are therefore warranted even for MVMR analyses.

Recall that invalid instruments exhibiting horizontal pleiotropy introduce heterogeneity in univariable MR causal effect estimates; the presence of pleiotropy in a single instrument can be tested by the $\hat{d}_i$-statistic 4.10 [126] and globally, over all the instruments, either by the Cochran's Q statistic 3.13 or the T_{HEIDI} statistic 4.11 (Subsection 6.3.1). In MVMR however, ascertaining the validity of instruments is not as straightforward—it is simply not possible to isolate genetic variants in the analysis because the number of instruments in the model needs to equal or exceed the number of exposures (Subsection 5.4.3). We can nevertheless leverage the experience from traditional MR towards a generalized approach for detecting pleiotropy/heterogeneity in multivariable settings.

In accordance with equation 5.9, the total causal effect of a valid instrument on the outcome should equal the sum of effects mediated by all the exposures. We can construct a test statistic similar to the $\hat{d}_i$-statistic 4.10 to test for differences between these two, assumed to be normally distributed with mean zero [3]:

$$\hat{d}_i = \hat{\beta}_{G_iY} - \sum_{j=1}^{k} \hat{\beta}_{G_iX_j}\hat{b}_j\,, \qquad \hat{d}_i \sim \mathcal{N}\left(0, \mathrm{Var}\left(\hat{d}_i\right)\right). \tag{6.9}$$

To be able to obtain the variance of the $\hat{d}_i$-statistic 6.9 with summary statistics alone, we could assume all individual estimates/terms in $\hat{d}_i$ to be independent and use the sum and product rule of variance [3]:

$$\mathrm{Var}\left(\hat{d}_i\right) = \mathrm{Var}\left(\hat{\beta}_{G_iY}\right) + \sum_{j=1}^{k}\left[\mathrm{Var}\left(\hat{\beta}_{G_iX_j}\right)\mathrm{Var}\left(\hat{b}_j\right) + b_j^2\mathrm{Var}\left(\hat{\beta}_{G_iX_j}\right) + \beta_{G_iX_j}^2\mathrm{Var}\left(\hat{b}_j\right)\right],$$

where b_j and $\beta_{G_iX_j}$ are the true effect of exposure X_j on Y and instrument G_i on X_j, respectively. Like in univariable MR, instruments deviating too much from the norm in terms of the $\hat{d}_i$-statistic 6.9 could be eliminated from the analysis. Furthermore, the sum of $\hat{d}_i$ over all the instruments allows to test for effect heterogeneity [137] just like Cochran's Q and T_{HEIDI} statistics.

The methods of causal inference invariably depend on assumptions [40], so a blind application of some methodology can easily lead astray. Investing into sensitivity analysis is thus always worthwhile in practice. Familiarity with the problem domain further helps to construct causal arguments and choose the most appropriate analysis approach. As we will see by examples in the following chapter, obtaining sound results in practice further benefits from triangulation of evidence from multiple orthogonal sources.

7. IDENTIFYING CAUSAL GENES IN PRACTICE

> Alice laughed: *There's no use trying*, she said; *one can't believe impossible things*.
>
> *I daresay you haven't had much practice*, said the Queen.
>
> Lewis Carroll *(Alice in Wonderland)*

In the previous chapters, we introduced in rather technical terms the principles of identifying causal relationships between gene expression and complex traits with statistical methods. It was to provide a foundation for understanding the scientific contribution of this dissertation. In the chapter at hand, we will focus on already published scientific articles that serve as the basis for this contribution. It entails both methodological innovations and uncovering novel putative causal relationships in practical applications. Since the articles have been reprinted at the end of this dissertation and could thus be visited for more details, we will be purposefully rather brief in our coverage here; we will merely show how to apply the theoretical framework that we have learned so far—and mold it whenever necessary—in such a way as to enable causal answers to specific questions in custom-tailored analyses.

7.1. Causal inference using small sample individual-level data (Ref. I)

In **Ref. I**, we set out to find causal links between C-reactive protein (CRP) and gene expression in the EGCUT cohort. Elevated levels of CRP in the blood are indicative of inflammation in the body [138]. While inflammation is the immune system's normal response to pathogens (e.g. viral, bacterial), tissue injury and other harmful stimuli, it can have detrimental effects in chronic form; lead to inflammatory diseases (e.g. CVD, T2D) and early mortality [139]. As a biomarker, CRP can be used in clinical practice to determine disease progress and measure treatment effectiveness [140–142]. Whether CRP shares any actual responsibility in the decline of immune function is an open question which is not well understood. Here, we tried to shed light on the matter.

7.1.1. Novel likelihood-based model selection approach to prioritize putative causal genes

A major obstacle to overcome in this study was the small sample size—there were only 491 individuals with overlapping gene expression and CRP measurements in the Estonian data. As we know, MR is not well suited to small samples due to limited power (Subsection 6.1.2). Instead, our approach (loosely inspired by [143]) was to exploit the rules of d-separation (Subsection 5.2.1) to establish possible

causal relationships between triplets (G,X,Y) of genetic variant G, gene expression trait X and CRP levels Y; evaluate respective model likelihoods, and prioritize best-fitting models. This boiled down to a multi-step analysis procedure.

First, we performed an association study between all gene expression traits and CRP; significant associations from such an analysis include causal relationships $X \to Y$ and $X \leftarrow Y$ but are also likely to emerge due to confounding (Subsection 5.3.1). Next, we conducted a *cis*-eQTL analysis with significant gene expression traits from the previous step, this time identifying causal $G \to X$ relationships. Together, the results of these two steps are sufficient for the following models:

(i) causal—genetic variant regulates CRP levels via gene expression mediation
(ii) colliding—genetic variant and CRP independently regulate gene expression
(iii) reverse—genetic variant regulates gene expression through CRP mediation
(iv) independent—genetic variant independently regulates both gene expression and CRP levels

For every triplet (G,X,Y), we thus had to calculate the likelihood of the four models and prioritize the likeliest one (more details in [1]). However, note that per our procedure, genetic variants were selected in the second step based on the strength of association with gene expression. This leads to selection bias favouring causal and colliding models over the others. We could thus simplify our analysis and determine a plausible causal direction between X and Y by the difference in AIC values of these two models in the EGCUT sample:

$$\Delta_{AIC} = \prod_{i=1}^{n} \underbrace{p_{causal}(G=g_i, X=x_i, Y=y_i)}_{p(g_i)p(x_i|g_i)p(y_i|x_i)} - \prod_{i=1}^{n} \underbrace{p_{colliding}(G=g_i, X=x_i, Y=y_i)}_{p(g_i)p(y_i)p(x_i|g_i,y_i)}, \tag{7.1}$$

where the following distributions were assumed:

$$\begin{aligned}
G &\sim \mathcal{B}\big(2, \text{frequency}(G)\big), \\
Y &\sim \mathcal{N}\big(\mathrm{E}Y, \mathrm{Var}(Y)\big), \\
X \mid G &\sim \mathcal{N}\big(\mathrm{E}X_G, \mathrm{Var}(X)\big), \\
Y \mid X &\sim \mathcal{N}\left(\mathrm{E}Y - \sqrt{\tfrac{\mathrm{Var}(Y)}{\mathrm{Var}(X)}}\rho_{XY}(X - \mathrm{E}X), \big(1-\rho_{XY}^2\big)\,\mathrm{Var}(Y)\right), \\
X \mid G, Y &\sim \mathcal{N}\left(\mathrm{E}X_G - \sqrt{\tfrac{\mathrm{Var}(X)}{\mathrm{Var}(Y)}}\rho_{XY}(Y - \mathrm{E}Y), \big(1-\rho_{XY}^2\big)\,\mathrm{Var}(X)\right).
\end{aligned}$$

In order to work with triplets that had considerable support for one or the other model, we required that $|\Delta_{AIC}| \geq 10$. Furthermore, causal triplets had to satisfy $G \not\perp\!\!\!\perp Y \wedge G \perp\!\!\!\perp Y \mid X$ while colliding triplets had to satisfy $G \perp\!\!\!\perp Y \wedge G \not\perp\!\!\!\perp Y \mid X$ (Subsection 5.2.1). Nevertheless, our likelihood-based approach was never designed to weed out all confounding, rather we attempted to prioritize candidate genes for additional investigations.

Altogether, the analysis above highlighted ten candidate genes which were brought forward to a two-sample MR analysis (Section 6.2). To account for trait complexities, increase power in the analysis and have independent sources of evidence, we used summary statistics from the biggest CRP GWAS [144] and eQTL study (then unpublished eQTLGen Consortium data [69]) available at the time. Among the ten candidates, only one involved a causal triplet model—the *FADS2* gene. To officially test for its causal effect on CRP levels ($H_0 : b_{XY} = 0$), we performed the SMR analysis 6.6 with the top instrument identified in the eQTL study; we did not reach statistical significance ($P = 0.0996$). For genes involved in colliding triplet models, we tested for the causal effect of CRP on expression ($H_0 : b_{YX} = 0$) using genetic variants associated with CRP levels ($P < 5 \times 10^{-8}$) as instruments in a multi-instrument MR setting with the IVW estimator 6.7, followed by MR-Egger regression to investigate the presence of horizontal pleiotropy (Subsection 6.3.1). A single gene, *CD59*, emerged as being putatively causally regulated by CRP levels ($\hat{b}_{YX} = 0.2$, $P = 0.0012$).

7.1.2. The importance of triangulation of causal evidence

Due to assumptions which are not testable in practice, isolated statistical analyses are not enough to warrant causal claims. Evidence is stronger if orthogonal methods and approaches lead to the same conclusions. In particular, the results should fit into the biological understanding of the mechanistic processes in question. Statistical results should ideally be validated in experimental settings, though this is rarely seen in computational analyses papers. In our study, we confirmed the causal link between CRP and *CD59* expression in blood with cell culture stimulation assays, adding considerable conviction to the finding.

While we could not gather enough support for a functional role of *FADS2* expression on CRP, it has been associated with lipid levels [145, 146] which in turn have been causally implicated in inflammatory processes and CRP [147]; *FADS2* could thus have a mediated indirect effect on CRP and we could have been underpowered to detect it. On the other hand, *CD59* is known to inhibit the formation of complement membrane attack complex which as part of the innate immune system induces the lysis and death of targeted cells during infections and inflammation [148]. The involvement of CRP in the regulation of *CD59* expression in blood thus suggests a negative feedback mechanism to control the immune response from damaging healthy blood cells. This could have implications in terms of understanding and controlling for low-grade chronic inflammation.

7.1.3. Other contributions to the field

The contributions of **Ref. I** to the wider scientific community are not only our methodological approach and novel results. As part of the study, the raw RNA-seq data of the EGCUT cohort was—for the first time—prepared, processed, analysed, and made publically available in the form of eQTL summary statistics. Fur-

thermore, note how we identified a link from a complex trait to gene expression, not vice versa. To the best of our knowledge, this had not been done before with computational methods. Causal links of such direction would have been—and still are!—difficult to uncover with MR due to relatively small sample sizes of gene expression data (Figure 6b). In addition to that, *trans*-eQTL summary statistics irrespective of the strength of association were not readily available at the time [149], likely due to overhead in data management (associations between all SNPs and gene expression traits are in tens or even hundreds of billions). We cooperated with the eQTLGen Consortium to access necessary summary statistics—based on the biggest eQTL study available, though unpublished at the time.

Finally, it is important to acknowledge that we endeavoured causal reasoning only after collecting causal evidence from multiple orthogonal sources. Experimental validation—made possible thanks to a direct collaboration with immunology Prof. Pärt Peterson from the University of Tartu—was particularly essential. All in all, our study shows the importance and significance of triangulation of evidence for causal reasoning.

7.2. Genes in 16p11.2 BP4-BP5 CNV region with a causal effect on age at menarche (Ref. II)

In **Ref. II**, we seeked to explain the effect of 16p11.2 600 kilobase (kb) BP4-BP5 CNV interval on reproductive traits. Copy number dosages of this interval have been linked to developmental disorders (e.g. autism), changes in brain structure, cognitive functioning, and extreme BMI [150–153]. Collectively, these traits imply a disrupted developmental process, characterizable by reproductive outcomes. Though pubertal timing has been associated with common SNPs in this interval in GWAS approaches [154], it has remained relatively understudied as part of the CNV phenotype. While CNVs are usually rare due to potentially very detrimental health outcomes, rearrangements of the 16p11.2 BP4-BP5 interval can be considered as fairly common with a frequency of 0.04% for deletion and 0.05% for reciprocal duplication [150]. For these reasons, careful investigation into the functional relationships at play here is of great clinical interest, specifically in terms of the genes responsible for the phenotypic tendencies. Of particular priority in this study was the timing of sexual development as measured by the beginning of puberty. Since this is better defined in women than in men, we concentrated primarily on age at menarche (AAM) as the trait of interest.

7.2.1. Puberty timing tracks with 16p11.2 BP4-BP5 dosage

Starting off with a UKBB-based association analysis, we showed for the first time that AAM tracks with copy number dosage of the 16p11.2 BP4-BP5 interval, representing a mirror effect; menstrual cycles of women carrying either the deletion or reciprocal duplication of this CNV region started about 1.5 years earlier (Wilcoxon $P = 0.001$) or later (Wilcoxon $P = 0.002$), respectively, compared to

the average of 12.9 years in normal controls. These results validated in mouse models, though effect directions were inverted compared to humans: female mice carrying either the deletion or duplication reached first ovulation about 5 days later (Wilcoxon $P = 4.8 \times 10^{-6}$) or 2 days earlier (Wilcoxon $P = 0.0021$) than their wild type littermates.

The 16p11.2 BP4-BP5 CNV interval encompasses and thus has consequences on 29 genes (Figure 10). To pinpoint genes causally modulating AAM, we naturally turned to MR. There is a caveat however—the physical proximity of many genes in such a small region results in genetic architectures that seem to be very similar due to LD (Figure 3). It becomes virtually impossible to differentiate valid instruments from eQTLs exhibiting horizontal pleiotropy, representing a problem of bias for MR analysis. We tackled the issue by complementing the traditional SMR+HEIDI approach (equations 6.6 and 4.11) [88] with the MVMR method 6.8 [136], hypothesizing that having all the BP4-BP5 genes in the model simultaneously—accounting for pleiotropic effects through these likely causal genes—could better help to deal with the bias.

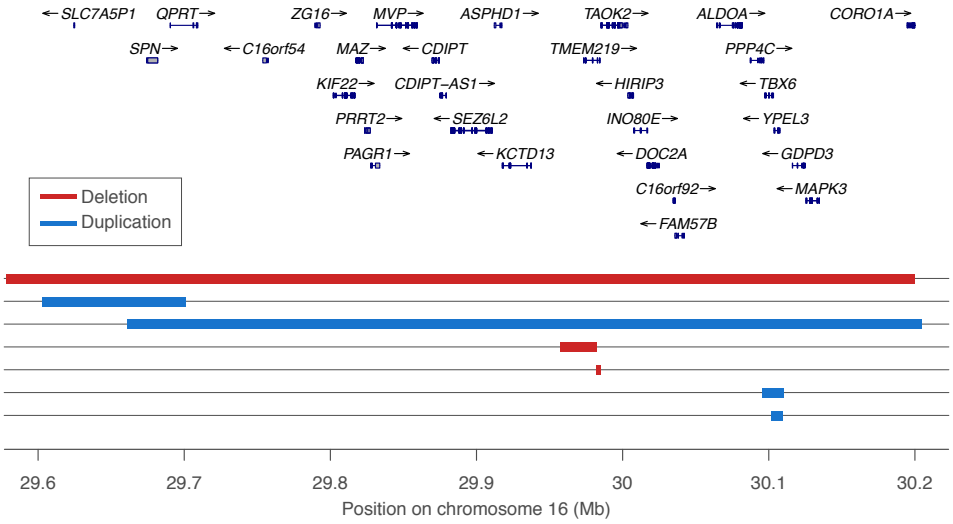


Figure 10: Loss of function deletions and copy gain duplications in gnomAD data [155] covering at least 1000 base pairs in the 16p11.2 BP4-BP5 CNV interval from $29.6-30.2$ Mb (bottom), and genes in the same region [82] (top). The bigger (≈ 600 kb) CNVs have consequences on all 29 genes in the region.

To maximize statistical power, we used summary statistics from the largest eQTL study (then unpublished eQTLGen Consortium data [69]) and GWAS on AAM [154] at the time. The eQTL instruments were identified in whole blood which is not necessarily the most informative tissue for AAM; however, sample sizes for more relevant tissues were very small ($N \approx 100$ for brain tissues in GTEx data [118]) and blood eQTLs have been shown to be good proxies for eQTLs of other tissues [156]. We used a P-value threshold 10^{-5} for the inclusion (corre-

sponds to $F \approx 20$ in equation 6.3) which guarantees strong instruments while not compromising on power like more stringent thresholds. In order to include only independent SNPs, we applied the stepwise fine-mapping strategy as implemented in GCTA software [83] (Subsection 4.1.1). Altogether, 12 out of the 29 genes had strong eQTLs in the eQTLGen data. These genes were tested for the causal effect on AAM with our MR approach, more than doubling the number of genes previously analyzed for this purpose [154]. The analysis supported a causal role in AAM for two genes: *INO80E* ($b_{XY} = 0.071$, $P = 9.3 \times 10^{-5}$) and *KCTD13* ($b_{XY} = -0.074$, $P = 9.7 \times 10^{-4}$).

7.2.2. External investigation into causal genes

Similarly to **Ref. I**, we embarked on an independent validation approach in experimental settings to validate the MR findings. Unlike CRP however, human AAM cannot be treated to a lab protocol, prompting us to turn to animal models instead. A useful indicator of the reproductive function is provided by the development of the gonadotropin-releasing hormone (GnRH) system which happens to be evolutionarily well conserved in vertebrates [157]. Zebrafish represents a particularly good model organism to study this due to its transparency in early development [158]. The animals are genetically modified such that GnRH neurons (of type 3 specifically) would express enhanced green fluorescent protein (EGFP) [157–159]. The development of the GnRH3 system in such transgenic (Tg) zebrafish can then easily be monitored in high-resolution using fluorescence imaging techniques [159]. We independently modulated the dosage of BP4-BP5 genes in the Tg:GnRH3:EGFP zebrafish line and measured the area of EGFP-positive cells in the dorsal aspect of larvae at 5 days post fertilization (dpf). Overexpression of a single gene, *ASPHD1*, significantly reduced the EGFP signal compared to controls (by 19%, linear regression $P < 0.0001$); other genes had no effect, not even *INO80E* nor *KCTD13* which were brought forward in the MR analysis. A similar reduction in EGFP signal at 5 dpf was seen in Tg:GnRH3:EGFP larvae depleted of *ASPHD1* transcripts with CRISPR/Cas9 genome editing technique [160] (by 13%, linear regression $P = 0.0031$).

We could not test for the causal effect of *ASPHD1* on AAM with MR due to data availability; *ASPHD1* is predominantly expressed in brain and pituitary gland [118], and did not have any eQTLs in our whole-blood based data. However, we hypothesized that MR-supported causal genes *INO80E* and *KCTD13* could interact with *ASPHD1* to exacerbate or mitigate its effect. Co-injecting *KCTD13* with *ASPHD1* indeed resulted in a significantly reduced EGFP signal compared to overexpression of just *ASPHD1* (by 14%, linear regression $P = 0.003$). Not too much is currently known about the biological function of *ASPHD1* but our results suggest it interacts with *KCTD13* and plays a role in the development of the reproductive system.

7.2.3. Summary of our contributions to the field

In conclusion, we showed an association with the copy number dosage of 16p11.2 BP4-BP5 interval and pubertal timing. More importantly, we applied an interdisciplinary approach of computational MR analyses and experiments on zebrafish models to tease out functionally responsible genes, identifying *ASPHD1* as a driver and *KCTD13* as a modifier of the reproductive phenotype. These are important findings for making sense of the underlying biological processes in sexual development and the aetiology of associated diseases.

The analysis we conducted herein confirms an important role for MR in establishing causal genes for complex traits but signifies the importance of orthogonal (experimental) approaches and tissue-specific scans. Causal analyses in tissues other than whole blood have so far remained scarce due to low power of MR and small sample sizes of available gene expression data. This is likely to change in the future with continuous collection of tissue-specific expression measurements by projects like GTEx [118], together with harnessing and systematic analysis of this data by initiatives such as the eQTL Catalogue [161]. Nevertheless, expression profiling in blood remains the most abundant and continues to represent the most viable option for MR studies in terms of providing raw power. As is evident from our analysis however, tissue-specific gene-trait relationships can be missed if blood is used as a proxy to other tissues. By integrating rare variant analyses, MR, and experiments on mouse and zebrafish models, we could overcome limitations of any single methodological approach. Our MR results fed directly into the design of lab experiments and helped to elucidate the interplay between gene expression and reproductive traits.

Our study in **Ref. II** represents a thorough and powerful investigation into the molecular mechanisms behind pubertal timing. Similarly to **Ref. I**, we back our claims using several sources of evidence, made possible thanks to a combined collaborative effort of many distinct researchers. Just the analyses with the highest relevance to the narrative of this dissertation have been described here. Refer to the paper at hand [2] for more details.

7.3. Mendelian randomization over the transcriptome (Refs. III, IV, V)

In **Ref. III**, we performed a systematic evaluation of MVMR for identifying causal genes over the entire transcriptome for 43 complex traits; in **Ref. IV**, we performed the analysis in men and women separately. We have already covered the theoretical suitability of MVMR for causal gene discovery in Section 6.4, corroborated by simulation studies [54, 137] and custom-tailored analyses (e.g. **Ref. II** [2]) alike. However, these examples demonstrating the promise of MVMR are sporadic and highly circumstance-specific. It is not clear whether the benefits of this methodology over competing strategies for gene prioritization (e.g.

fine-mapping, single-gene MR) are generalizable to less defined situations and hypothesis-free scans. Here, we investigated these matters in detail.

7.3.1. Practical considerations of transcriptome wide analysis

Applying MVMR to estimate the causal effect of an arbitrary gene on an outcome raises the question of how to determine accompanying genes and respective instruments to include in the model. Since eQTLs are more likely to manifest direct causal effects in *cis* as opposed to act on distant genes [135], a general strategy could be to incorporate in the model all the genes in a genomic region of interest. In our approach, called transcriptome wide Mendelian randomization (TWMR), this is implemented as follows:

1. For a focal gene X_f, let its instruments be eQTLs that pass a pre-specified P-value threshold in an eQTL study: $\mathcal{G}_f = \{G \in \mathcal{G} \mid P_{G,X_f} \leq \text{threshold}\}$, where $\mathcal{G}$ denotes for the set of all SNPs.
2. From the set of all gene expression exposures $\mathcal{X}$, find genes that share eQTLs with the focal gene: $\mathbf{X} = (X \in \mathcal{X} \mid P_{G,X} \leq \text{threshold} \wedge G \in \mathcal{G}_f)$.
3. Define instruments as eQTLs of all the previously established exposures, $\mathbf{G} = \{G \in \mathcal{G} \mid P_{G,X} \leq \text{threshold}, X \in \mathbf{X}\}$.

Completing the steps above already facilitates the MVMR approach 6.8 which naturally allows for LD between instruments. Nevertheless, we pruned eQTLs to be nearly independent in order to account for the weak instrument bias of MR (Subsection 6.1.1). Only then did we apply formula 6.8 to estimate the causal effect of the focal gene X_f on the outcome, correcting for heterogeneity in effect estimates by eliminating outlying instruments as per the $\hat{d}_i$-statistic 6.9.

We used summary statistics from the largest eQTL and GWAS meta-analyses at the time to run the algorithm. Like in **Ref. I** and **Ref. II**, eQTLs once again originated from data by the eQTLGen Consortium [69]. As reasoned previously, this choice of data maximizes statistical power, even though blood may not necessarily be the causal tissue for some traits.

7.3.2. Improving upon existing approaches to implicate novel causal gene-trait relationships (Ref. III)

Among 19251 genes and 43 traits, we found 3913 significant associations after correcting for multiple tests (P-value $< 0.05/16000$, where 16000 corresponds to the approximate number of genes tested for each trait); 36% of these relationships were not previously prioritized by GWAS and fine-mapping approaches. To study the reasons behind this apparent discrepancy, we conducted both GWAS and TWMR analyses with BMI in randomly chosen subsets of various different sizes of the UKBB data, defining GWAS hits as all genes within 500 kb region from significant SNPs. As could be expected, the number of BMI-related genes increased linearly with sample size irrespective of the methodology; more importantly however, genes missed by GWAS but detected by TWMR in smaller sample

sizes tended to eventually be confirmed by GWAS in bigger sample sizes. This provides evidence for the superiority of TWMR in terms of statistical power.

In total, 848 of the TWMR-recommended genes were causal for several traits, indicating widespread pleiotropy like other studies before us [41, 130]. Dealing with pleiotropy is crucial to minimize biases in causal effect estimation. To further our understanding in how the MVMR approach fares compared to single-gene MR in accurately estimating causal effects in realistic settings, we designed a simulation study. We used both methods to estimate the causal effect in simulation models with varying number of genes (e.g. 3 or 5), SNPs (e.g. 15 or 30), and degrees of pleiotropy that each SNP exhibits on gene expression (i.e. the number of genes it influences on average). MVMR was universally superior in terms of mean squared error (MSE) between true and estimated causal effects. The single-gene MR suffered from inflated T1E rate (upwards 20%) while MVMR did not exceed the 5% nominal level. Importantly, the benefits of lower MSE and T1E did not come at the expense of power—both methods fared comparably in that regard. In brief, MVMR for causal inference in our simulations was considerably less affected by pleiotropy than single-gene MR while not compromising on power.

One of our many novel findings of causal gene-trait relationships is the mediating role of intellectual impairment-associated *BSCL2* [162] in the culmination of educational attainment. However, as well as implicating functional genes in regions not previously prioritized by GWAS for corresponding traits, we could also reassign causal genes in regions already flagged as trait-associated. For example, we found a causal link between a short stature-associated *CRIPT* [163] and height in a locus where another gene, *SOCS5*, was previously prioritized instead [164]. Altogether, we observed many examples where the top GWAS SNP did not lead to a TWMR-implicated gene, indicating that the physically closest gene to the most trait-significant genetic variant in the region need not always be causal, like shown elsewhere [165, 166].

The results and examples in **Ref. III** really are plentiful; we restated only the more interesting findings here. Refer to the actual paper (reprinted at the end of the dissertation) and its supplementary [3] for further exploration.

7.3.3. Sex-specific effects (Ref. IV)

The results that we have covered so far were found by analyzing entire populations without any regard to obvious discrepancies between some stratums of individuals. Most notably, men and women present differences in characteristics of nearly all complex traits and diseases: in susceptibility (incidence and prevalence), age of onset, progression, severity, etc. [167]. Not accounting for that can lead to missed discoveries and incorrect estimates due to effect dilution. The main culprit for the apparent disregard to sex-specificity of complex traits in association studies lies in the difficulty in detecting potentially small differences in effect sizes between strata. After all, genetic variants exhibit at most only a tiny influence on

trait variability by the polygenic model [20], requiring large sample sizes to discover as is. Sex-stratified analyses need to make do with roughly half the available samples to detect differences in effects which can be—by design—even smaller. Until recently, sex-specific analyses have simply been underpowered to even attempt performing. It is therefore not surprising that drugs have been developed and prescribed following a one-size-fits-all paradigm, leading to variable efficacy between sexes, and adverse reactions [168]. In order to truly advance precision medicine, it is paramount to account for sex-specificity in complex traits and diseases, and understand the mechanisms responsible for phenotypic differences between sexes. Despite recognizing the difficulty of the task, we set out to explore these problems in **Ref. IV**.

Our undertaking was motivated by the hypothesis that variation in phenotype values between men and women is at least partly driven by differences in gene expression regulation. To enable inference to that end, we first identified sex-specific eQTLs from 1 928 women and 1 519 men in whole blood RNA-seq data collected by the BIOS Consortium (https://www.bbmri.nl/acquisition-use-analyze/bios/). Using a t-test, we found 18 genes with sex-specific eQTLs. We performed a UKBB-based phenome-wide association study with lead eQTLs of these genes and over 700 traits, identifying some associations with morphological and hematological traits; however, these associations did not replicate in sex-stratified analysis of corresponding traits, indicating no enrichment of sex-specific GWAS hits among sex-specific eQTLs. To also approach from the other side, we started off with sex-specific association studies of two traits with considerable differences between men and women—WHR and testosterone levels—but did not see any enrichment of sex-specific eQTLs among sex-specific GWAS findings either.

While it is still too early to say how much of sex-specificity in traits is down to sex-specificity in gene expression regulation—we show by simulations that the power to test for this is simply too low, requiring millions of samples—a sex-specific TWMR analysis on the traits above (WHR and testosterone levels) turned out to be more fruitful, uncovering several sex-specific causal genes, e.g. *IFT27* with testosterone levels in men and *CCDC92* with WHR in women. Importantly, a negative control TWMR analysis on educational attainment did not show any sex-specific associations, providing some validation for our approach. Thus TWMR can help to identify putative mechanisms underlying sex differences in complex traits, even if traditional GWAS approaches are unable to do so.

7.3.4. Reverse causation: from traits to expression (Ref. I, V)

Studies looking for causal relationships between gene expression and complex traits have predominantly focused on only one causal direction, gene $\rightarrow$ trait. There are two major reasons why trait $\rightarrow$ gene relationships have not been investigated as much. First, it is arguably more useful to identify causal genes underlying diseases because these could be used to develop drugs for the cure or intervention

of said diseases. Second, exploring trait influences on the transcriptome requires *trans*-eQTLs which have not been widely available until recently. Our analysis in **Ref. I** [1]—remember that we reported a causal effect of circulating CRP levels on *CD59* expression—represents one of the few and earliest exceptions in this regard (Subsection 7.1.3). As part of the study, we curiously noticed that observed gene-trait correlations in EGCUT data do not overlap with causal gene → trait associations identified with TWMR in **Ref. III** (Figure 11). Due to data availability, we analyzed a subset of 22 out of all the complex traits investigated in **Ref. III**, including blood cell counts, hemodynamic parameters, anthropometry measurements, lipid levels, and diseases with at least 20 cases. All quantitative trait values were measured from blood samples where gene expression was quantified, or on the day when samples were taken. For diseases, individuals were classified as cases if they had been diagnosed with the disease by this day, controls otherwise.

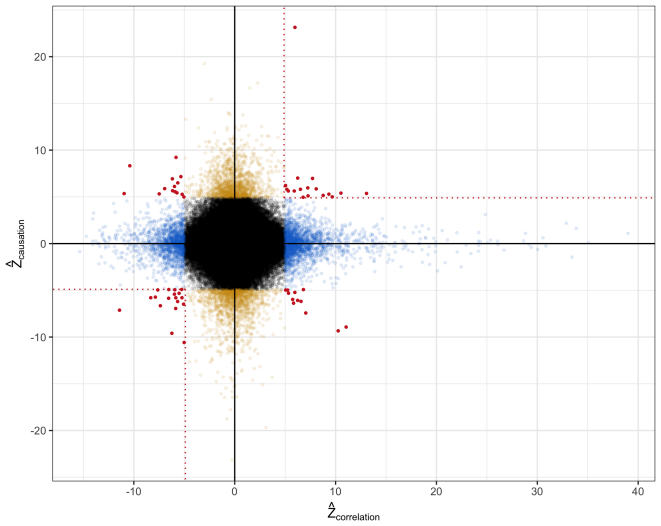


Figure 11: Scatter plot of $\hat{Z}$-statistics from the EGCUT-based correlation analysis (**Ref. I**) on x-axis and $\hat{Z}$-statistics from TWMR analysis (**Ref. III**) on y-axis. Each point represents a gene expression-outcome pair. Blue and yellow points correspond to just correlated xor causally associated relationships, respectively, while red points indicate significance in both analyses (note that only those in the first and third quadrants exhibit consistent effect direction).

While correlations can emerge due to confounding and should thus not be used to infer causal relationships by theory (Subsection 5.3.1), we would still have expected some enrichment of causality. In **Ref. V**, we theorize on the possibility that gene-trait correlations could be better described by the reverse direction: causal trait → gene associations. This would be consistent with a recent large-scale transcriptome study which found that only a small fraction (4%) of *trans*-eQTL effects can be explained by mediation via genes in *cis*, indicating widespread environmental causes in gene expression regulation [69]. Research on this is ongoing.

7.3.5. Our contributions to the field

Here, we show convincing evidence that MVMR has more power to identify causal genes compared to GWAS and fine-mapping approaches. Further, it corrects for some of the pleiotropy-induced bias in the causal effect estimates while having a reduced T1E rate compared to single-gene MR. Applying MVMR in our transcriptome wide scan lead to the identification of many putative causal genes for various complex traits.

The plethora of causal gene-trait relationships that we uncovered in **Ref. III** gives predence to the claim that gene expression acts as a mediator between genetic variants and complex traits. In addition, note that eQTLs—used as instruments in MR—by definition lead to discoveries of expression-induced associations with complex traits; genetic variants responsible for changes in the protein structure of genes are simply not targeted by eQTL studies. In fact, the TWMR approach correctly did not uncover known gene-trait associations induced by coding variants. Therefore, our findings are indicative of extensive regulatory effects acting on complex traits on the gene level. We also show widespread sharing of genetic architecture between traits, corroborated by other studies [41, 130]. Insight from our analyses coupled with the notion that just computational tools were used to prioritize causal relationships is promising for the pharmaceutical industry concerned with the design and development of candidate disease interventions.

In an attempt to provide insight for precision medicine approaches, we investigated in **Ref. IV** whether sex-specificity in complex traits can be attributed to sex-specificity in gene expression regulation. We showed by power calculations that available gene expression data is currently not abundant enough to answer this question with traditional GWAS approaches, yet sex-specific TWMR can be applied to yield sex-specific causal genes.

Even though correlations should not be used to inform causal links due to confounding, true causality implies correlation (Subsection 5.3.1). However, we did not observe any enrichment of causal gene $\rightarrow$ trait relationships among significant gene expression-trait correlations in **Ref. V**. Instead, we argued that causal trait $\rightarrow$ gene relationships might be more prevalent among such correlations.

Like in **Ref. I** and **Ref. II**, our results in **Refs. III-V** are based on expression data originating from whole blood. Analyses in other tissues have the potential to uncover tissue-specific effects and thus would likely result in the identification of even more functional relationships. Given our understanding about the superior power of TWMR over GWAS for implicating causal genes, this leads us to hypothesize that future efforts in teasing out yet more causal links between genes and complex traits could have bigger benefits from concentrating on tissue-specific analysis, instead of increasing GWAS sample size. Like argued in Subsection 7.2.3, the feasibility and meaningfulness of this endeavour is likely to increase in the near future when additional gene expression data from non-blood tissues becomes more abundant and widely available.

8. CONCLUSION

> I almost wish I hadn't gone down that rabbit-hole — and yet — and yet — it's rather curious, you know, this sort of life!
>
> Lewis Carroll's Alice
> *(Alice in Wonderland)*

The material presented in this dissertation is—of necessity—a mix from several different fields, incorporating statistical methods from regression to meta-analysis (Chapter 3), association-based approaches for gene prioritization (Chapter 4), the formal concept of causality by Pearl (Chapter 5), the method of instrumental variables in econometrics (Section 5.4), how all these intertwine to enable MR together with sensitivity analyses (Chapter 6) for the purposes of causal inference in human genetics, and applied research (Chapter 7). The methods for disentangling causal relationships from observational data come in slightly different flavours but all are based on the same causal theory and—as we have hopefully managed to convince—are fundamentally very similar. This claim holds even for the methods that we did not explicitly cover in this dissertation, e.g. the popular regression-based approach for mediation analysis by Baron and Kenny [107] followed by the Sobel test [169], the causal inference test [108], network-based causal analysis methods [109], or structural equation models in general [40] (Section 5.2). Making causal inference from these methods is built on very strong assumptions requiring preemptive knowledge about underlying structural relationships between variables [110]. In contrast, MR, even though it comes with its own assumptions, is more robust by allowing for unmeasured confounders of the exposure-outcome relationship. This makes it applicable in more realistic settings, which is exactly the reason why we have given it the most attention in this dissertation.

8.1. In terms of teaching potential

Along with covering some of my more important research contributions over the past years (published in scientific journals), the core objective of this dissertation was to bring together connected theories and methods of mathematics and statistics in the field of causal inference and present them uniformly in the context of human genetics. There is an understandable tendency among scientific communities dealing with the intricacies of different concepts and methodologies to internalize or take for granted certain knowledge, and develop an intrinsic way of expressing related material. Addressing similar concepts and methodologies in an inharmonious manner can give rise to misconceptions and confusion over their utility and applicability in varying scenarios. The organisation and structuring of this dissertation was strongly influenced by my desire to avoid this and provide a coherent, theoretically sound treatment of inherently linked topics.

The analysis of causality is becoming very popular in human genetics due to increasing sample sizes; especially MR has gained a lot of traction over the last years. This further necessitates the placement of methods in context relative to each other, in terms of causality and what they can or cannot be used for. While scientific papers often assume existing knowledge, this dissertation aims to cover all the necessary material from the ground up and somewhat demystify the concept of causality in the process. My hope is that it will be useful for the new generation of (early career) researchers—in Estonia but not limited to—interested in the field of statistical genetics. The material presented here could, in the future, serve as a graduate level course in statistical genetics, either given by me or somebody else. For that reason, I tried my best to write this dissertation in a review-like tutorial-like manner, while not compromising too much on theoretical detail.

8.2. In terms of scientific research

It should be clear from the evidence presented in this dissertation that causality cannot be determined easily based on statistical reasoning alone. Sensitivity analyses can often help to highlight the violations of method assumptions and should thus be attempted along with formal causal analysis. Even so, implications to results and findings from untestable assumptions should be considered. Triangulation of causal evidence from orthogonal sources of information is especially important for increasing the trustworthiness of causal reasoning (**Refs. I** and **II**). Like is usually the case, the best approach to solving any problem depends on the problem domain; deep familiarity and understanding of it not only helps to choose the most appropriate methodology, it also facilitates arguments based on non-statistical evidence. While statistical methods are tremendously promising in speeding up causal discoveries in bulk (**Ref. III**), the final arbiter of causal results will remain to be experimental evidence [30].

An important reason for the increasing popularity of MR is that it can be performed using just summary statistics. For that reason, public databases of eQTLs (e.g. eQTLGen data [69] and the eQTL Catalogue [161]) and GWAS associations (e.g. the UKBB-based GeneATLAS [120], LD Hub [170], and PhenoScanner [171]) make MR applicable over a wide range of phenotypes, both molecular and clinical. The platform MR-Base [172] integrates several of such resources and even implements traditional MR methods together with sensitivity analyses to facilitate causal inference phenome-wide. Simply put, it has never been easier to attempt causal analyses in human genetics. All this means an unprecedented access into exploring the aetiology of disease and prioritizing drug targets for clinical trials [37, 173]. Indeed, phenome-wide MR studies have already been utilized for new discoveries of causal relationships between traits [174]. I have followed suit in my research, tackling functional genomics by developing and applying causal inference methods to identify causal genes for complex traits and diseases.

In **Ref. I**, we applied a likelihood-based causal analysis framework together

with MR and lab experiments to elucidate causes and consequences between CRP and gene expression. We uncovered a causal effect of CRP on *CD59* expression in blood, leading us to theorize about its protective function regarding healthy blood cells during human immune response. Causal links of such direction (from complex trait to gene expression) are hard to detect with computational tools—though we theorize in **Ref. V** that simple gene-trait correlations might be enriched of such relationships—due to scarcity of gene expression data. Our success in that regard comes down to both integration of different data and methods, and working as part of an interdisciplinary team of scientists with varying skills and expertise, leading to orthogonal sources of causal evidence and triangulation thereof.

In **Ref. II**, we shed light on functional genes mediating sexual development in the 16p11.2 BP4-BP5 CNV interval, pinpointing *ASPHD1* as a main driver and *KCTD13* as a modifier of the CNV phenotype. Starting off with a computational analysis, we decided for a MVMR approach to prioritize a list of AAM-relevant genes, in turn feeding these into experiments with zebrafish models. Coupled with rare variant association analyses on humans and mouse models, we provided unprecedented insight into development of reproductive traits and aetiology of associated diseases. These findings can once again (similarly to **Ref. I**) be attributed to our integrative interdisciplinary approach involving expertise in different scientific disciplines.

In **Ref. III** and **Ref. IV**, we applied MVMR over the entire transcriptome on a wide variety of complex traits—both over all individuals and stratified by sex—to identify actionable genes that could be brought forward to further experiments. Our method, TWMR, discovered causal genes with superior power compared to GWAS and fine-mapping approaches, and with a reduced T1E rate compared to single-gene MR. We uncovered a plethora of putative causal gene-trait relationships. However, as small sample sizes of gene expression data rendered sex-specific analyses largely unfruitful, it still remains to be seen in what portion can sex-specificity in phenotypes be attributed to differences in the regulatory level.

In conclusion, computational approaches represent a powerful means to gene prioritization and are already being used for selecting drug targets for RCTs. However, it should be kept in mind that causal analysis on observational data comes with a set of assumptions that necessitate understanding of methods and careful interpretation of results. Limitations due to data availability have resulted in most of the studies, including ours, to be conducted on whole blood gene expression measurements. This is despite our understanding that blood may not be the causal tissue for traits, and gene expression may not be a good proxy for gene products (proteins) that drugs would target.

8.2.1. Future directions

The methods introduced in this dissertation are directly applicable to omics layers beyond transcriptomics, such as proteomics. In the future, protein assays

and tissue-specific gene expression quantification are likely to propel analyses in these directions and further causal discoveries. The primary reasoning behind this notion is the continued acquisition of data by biobanks and research initiatives worldwide. With sufficient data, MR would become feasible in molecular phenotypes where sample size is currently the limiting factor (Subsection 6.1.2). In this aspect, the advancement of scientific knowledge on causal relationships depends on technological developments and respective costs in harnessing the necessary data from the biological material of biobank participants. Recent MR analyses with a subset of the proteome on a phenome-wide scale have already yielded new insight into disease processes and drug target prioritization [175].

Yet more insight into causal processes is likely to stem from statistical method development. Traditional MR analyses are isolated to specific exposure and outcome traits at a time (Section 6.1), unable to capture more sophisticated relationships between several variables. However, biological systems can rather be thought of as complex networks of genetic and environmental factors [39]. To uncover the structure of these networks, known and verified causal relationships should either be incorporated into new investigations step-by-step, or confounding robust methods should be developed that permit to build causal networks from the ground up. These involve mediation analyses able to distinguish between direct and indirect causal effects. Integration of different data types is necessary for all such approaches. Initial frameworks for elucidating causal networks incorporating all the available phenotype information of national biobanks have recently been proposed [176].

The major limiting factor of causal enquiries is the inability to verify core method assumptions (Section 5.4). Computational results need to undergo biological validation but this represents a bottleneck. In my opinion, the capacity to prioritize causal discoveries could be greatly enhanced if methods could be benchmarked and tested against reliable standards on simulated data. For this purpose, the functioning of biological systems would need to be adequately captured by simulation procedures—accurately capturing genetic architectures of complex traits (Subsection 2.2.2) would suffice. Quantifying the reliability of causal results obtained through statistical and computational means would facilitate focused method development and subsequent discovery of reliable insight into disease processes.

Appendix A. DERIVATION OF THE MULTIVARIABLE MENDELIAN RANDOMIZATION STANDARD ERROR

Consider the multivariable causal effect estimator 6.8:

$$\hat{b} = (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \hat{\Gamma}' C^{-1} \hat{\gamma},$$

where $\hat{\Gamma} : m \times k$ is a matrix of standardized eQTL effect sizes such that $\hat{\Gamma}_{ij} = \hat{\beta}_{G_i X_j}$ is the standardized effect of instrument G_i on exposure X_j; $C : m \times m$ is an LD-matrix between SNPs, and $\hat{\gamma} : m \times 1$ is a vector of standardized trait effect sizes such that $\hat{\gamma}_i = \hat{\beta}_{G_i Y}$ is the standardized effect of G_i on the outcome Y. Estimator $\hat{b}$ is essentially a function $f : \mathbb{R}^{m(k+1)} \to \mathbb{R}^k$, where k is the number of risk factors and m is the number of SNPs. Define the following:

$$\hat{\beta} = \begin{pmatrix} \mathrm{vec}(\hat{\Gamma}) \\ \hat{\gamma} \end{pmatrix},$$
$$\sigma_{\hat{\beta}} = \left(\sqrt{\mathrm{Var}(\hat{\beta}_1)}, \sqrt{\mathrm{Var}(\hat{\beta}_2)}, \ldots, \sqrt{\mathrm{Var}(\hat{\beta}_{m(k+1)})} \right)',$$

such that $\hat{\beta}$ is the vector of estimated effects of instruments on exposures and outcome—let $\beta = \mathrm{E}(\hat{\beta})$ denote the true effects—and $\sigma_{\hat{\beta}}$ is the corresponding vector of standard errors.

The Delta method gives us:

$$f(\hat{\beta}) \approx f(\beta) + J(\beta)(\hat{\beta} - \beta), \tag{A.1}$$

where $J : k \times m(k+1)$ is a Jacobian matrix. We can use the fact that $\mathrm{E}(\hat{\beta}) = \beta$ to approximate the average causal effect as

$$\mathrm{E}(\hat{b}) = \mathrm{E}(f(\hat{\beta})) \approx f(\beta). \tag{A.2}$$

In turn, the variance can be approximated using A.2 and A.1:

$$\begin{aligned}
\mathrm{Var}(\hat{b}) &= \mathrm{Var}(f(\hat{\beta})) \\
&= \mathrm{E}\left(\left(f(\hat{\beta}) - \mathrm{E}(f(\hat{\beta})) \right) \left(f(\hat{\beta}) - \mathrm{E}(f(\hat{\beta})) \right)' \right) \\
&\approx \mathrm{E}\left(\left(f(\hat{\beta}) - f(\beta) \right) \left(f(\hat{\beta}) - f(\beta) \right)' \right) \\
&\approx J(\beta) \mathrm{E}\left(\left(\hat{\beta} - \beta \right) \left(\hat{\beta} - \beta \right)' \right) J'(\beta) \\
&= J(\beta) \Sigma J'(\beta).
\end{aligned} \tag{A.3}$$

The covariance matrix of $\hat{\beta}$ under fixed genotypes is $\Sigma = \sigma_{\hat{\beta}} \sigma'_{\hat{\beta}} \odot (R \otimes C)$, where R is the correlation matrix of risk factors and outcome. Taking $R = I_{k+1}$ for simplicity allows to estimate Σ using just summary statistics and a genotype reference.

The Jacobian matrix J is the matrix of all first-order partial derivatives of a function, in our case f. Thus we need to take derivatives to find its value at β. However, true effects are unknown in reality. For that reason, we will evaluate J at the estimated effects $\hat{\beta}$ instead. For simplicity, we will omit the argument altogether and write J in terms of the causal effect estimator $\hat{b}$ which conveniently already incorporates all the individual estimates in the form of the eQTL effect size matrix $\hat{\Gamma}$ and GWAS effect size vector $\hat{\gamma}$:

$$J = \left(\frac{\partial \hat{b}}{\partial \hat{\Gamma}} \,\middle|\, \frac{\partial \hat{b}}{\partial \hat{\gamma}} \right), \tag{A.4}$$

where the vertical line is used to separate the two blocks of the Jacobian matrix J. First, we will find the derivative over $\hat{\Gamma}$:

$$\begin{aligned}
\frac{\partial \hat{b}}{\partial \hat{\Gamma}} &= \frac{\partial (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \hat{\Gamma}' C^{-1} \hat{\gamma}}{\partial \hat{\Gamma}} \\
&= \frac{\partial (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \hat{\Gamma}' C^{-1} \hat{\gamma}}{\partial (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \hat{\Gamma}'} \frac{\partial (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \hat{\Gamma}'}{\partial \hat{\Gamma}} \\
&= \left((C^{-1} \hat{\gamma})' \otimes \mathbf{1}_k \right) \frac{\partial (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \hat{\Gamma}'}{\partial \hat{\Gamma}}.
\end{aligned}$$

In order to facilitate the derivation, let us first define variables $S := (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1}$ and $T := \hat{\Gamma}'$. Now we can express

$$\begin{aligned}
\frac{\partial (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \hat{\Gamma}'}{\partial \hat{\Gamma}} &= \frac{\partial (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} T}{\partial (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1}} \bigg|_{T=const} \frac{\partial (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1}}{\partial \hat{\Gamma}} + \frac{\partial S \hat{\Gamma}'}{\partial \hat{\Gamma}'} \bigg|_{S=const} \frac{\partial \hat{\Gamma}'}{\partial \hat{\Gamma}} \\
&= \left(\hat{\Gamma} \otimes \mathbf{1}_k \right) \frac{\partial (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1}}{\partial \hat{\Gamma}} + \left(\mathbf{1}_m \otimes (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \right) P_{m,k},
\end{aligned}$$

where $P_{m,k} : mk \times mk$ is a commutation matrix (used to transform a vectorized matrix to its vectorized transpose). We have

$$\begin{aligned}
\frac{\partial (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1}}{\partial \hat{\Gamma}} &= \frac{\partial (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1}}{\partial \hat{\Gamma}' C^{-1} \hat{\Gamma}} \frac{\partial \hat{\Gamma}' C^{-1} \hat{\Gamma}}{\partial \hat{\Gamma}} \\
&= \left(-(\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \otimes (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \right) \frac{\partial \hat{\Gamma}' C^{-1} \hat{\Gamma}}{\partial \hat{\Gamma}}.
\end{aligned}$$

Once again, let us define variables $S := \hat{\Gamma}'$ and $T := C^{-1} \hat{\Gamma}$ to facilitate the subsequent derivation. Then

$$\begin{aligned}
\frac{\partial \hat{\Gamma}' C^{-1} \hat{\Gamma}}{\partial \hat{\Gamma}} &= \frac{\partial \hat{\Gamma}' T}{\partial \hat{\Gamma}'} \bigg|_{T=const} \frac{\partial \hat{\Gamma}'}{\partial \hat{\Gamma}} + \frac{\partial S C^{-1} \hat{\Gamma}}{\partial C^{-1} \hat{\Gamma}} \bigg|_{S=const} \frac{\partial C^{-1} \hat{\Gamma}}{\partial \hat{\Gamma}} \\
&= \left((C^{-1} \hat{\Gamma})' \otimes \mathbf{1}_k \right) P_{m,k} + (\mathbf{1}_k \otimes \hat{\Gamma}')(\mathbf{1}_k \otimes C^{-1}).
\end{aligned}$$

Putting everything above together leads to a rather complicated expression. We will simplify to get rid of the identity and commutation matrices, and eventually reach a somewhat more manageable solution:

$$
\begin{aligned}
\frac{\partial \hat{b}}{\partial \hat{\Gamma}} =& \frac{\partial (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \hat{\Gamma}' C^{-1} \hat{\gamma}}{\partial \hat{\Gamma}} \\
=& \left((C^{-1}\hat{\gamma})' \otimes \mathbf{1}_k\right) \Big[(\hat{\Gamma} \otimes \mathbf{1}_k) \left(-(\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \otimes (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \right) \cdot \\
& \cdot \Big(\left((C^{-1}\hat{\Gamma})' \otimes \mathbf{1}_k \right) P_{m,k} + (\mathbf{1}_k \otimes \hat{\Gamma}')(\mathbf{1}_k \otimes C^{-1}) \Big) + \\
& + (\mathbf{1}_m \otimes (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1}) P_{m,k} \Big] \\
=& \left((C^{-1}\hat{\gamma})' \otimes \mathbf{1}_k\right) \Big[\left(-\hat{\Gamma} (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \otimes (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \right) \cdot \\
& \cdot \Big(\left((C^{-1}\hat{\Gamma})' \otimes \mathbf{1}_k \right) P_{m,k} + (\mathbf{1}_K \otimes \hat{\Gamma}')(\mathbf{1}_k \otimes C^{-1}) \Big) + \\
& + \left(\mathbf{1}_m \otimes (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \right) P_{m,k} \Big] \\
=& \left((C^{-1}\hat{\gamma})' \otimes \mathbf{1}_k\right) \left(-\hat{\Gamma} (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \otimes (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \right) \Big(\left((C^{-1}\hat{\Gamma})' \otimes \mathbf{1}_k \right) P_{m,k} \Big) + \\
& + \left((C^{-1}\hat{\gamma})' \otimes \mathbf{1}_K\right) \left(-\hat{\Gamma} (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \otimes (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \right) \Big(\left(\mathbf{1}_k \otimes (C^{-1}\hat{\Gamma})' \right) \mathbf{1}_{mk} \Big) + \\
& + \left((C^{-1}\hat{\gamma})' \otimes \mathbf{1}_k\right) (\mathbf{1}_m \otimes (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1}) P_{m,k} \\
=& \left(-\hat{\gamma}' C^{-1} \hat{\Gamma} (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \otimes (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \right) \Big(\left((C^{-1}\hat{\Gamma})' \otimes \mathbf{1}_k \right) P_{m,k} \Big) + \\
& + \left(-\hat{\gamma}' C^{-1} \hat{\Gamma} (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \otimes (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \right) \Big(\left(\mathbf{1}_k \otimes (C^{-1}\hat{\Gamma})' \right) \mathbf{1}_{mk} \Big) + \\
& + \left((C^{-1}\hat{\gamma})' \otimes (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \right) P_{m,k} \\
=& \left(-\hat{\gamma}' C^{-1} \hat{\Gamma} (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \hat{\Gamma}' C^{-1} \otimes (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \right) P_{m,k} + \\
& + \left(-\hat{\gamma}' C^{-1} \hat{\Gamma} (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \otimes (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \hat{\Gamma}' C^{-1} \right) + \\
& + \left((C^{-1}\hat{\gamma})' \otimes (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \right) P_{m,k} \\
=& P_{k,1} \left(-\hat{\gamma}' C^{-1} \hat{\Gamma} (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \hat{\Gamma}' C^{-1} \otimes (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \right) P_{m,k} + \\
& + \left(-\hat{\gamma}' C^{-1} \hat{\Gamma} (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \otimes (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \hat{\Gamma}' C^{-1} \right) + \\
& + P_{k,1} \left((\hat{\gamma}' C^{-1}) \otimes (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \right) P_{m,k} \\
=& \left((\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \otimes \left(-\hat{\gamma}' C^{-1} \hat{\Gamma} (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \hat{\Gamma}' C^{-1} \right) \right) + \\
& + \left(-\hat{\gamma}' C^{-1} \hat{\Gamma} (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \otimes (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \hat{\Gamma}' C^{-1} \right) + \\
& + \left((\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \otimes (\hat{\gamma}' C^{-1}) \right) .
\end{aligned}
\tag{A.5}
$$

We used the following properties to get (A.5):

- $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$
- $A(B+C) = AB + AC$
- $A(BC) = (AB)C$
- $P_{k,1} = \mathbf{1}_k$
- $P_{k,1}(A \otimes B)P_{m,k} = B \otimes A$, where $A : 1 \times m$ and $B : k \times k$
- $A \otimes (B+C) = (A \otimes B) + (A \otimes C)$

To complete the derivation, we will also have to find the derivative over $\hat{\gamma}$. Luckily, this is less of a mouthful:

$$\frac{\partial \hat{b}}{\partial \hat{\gamma}} = \frac{\partial (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \hat{\Gamma}' C^{-1} \hat{\gamma}}{\partial \hat{\gamma}} = (\hat{\Gamma}' C^{-1} \hat{\Gamma})^{-1} \hat{\Gamma}' C^{-1} . \tag{A.6}$$

Substituting (A.5) and (A.6) into (A.4) and using the latter to estimate $J(\beta)$ in (A.3) gives us everything we need to estimate $\mathrm{Var}(\hat{b})$, the variance-covariance matrix of the multivariable Mendelian randomization causal effect vector.

BIBLIOGRAPHY

[1] Kaido Lepik, Tarmo Annilo, Viktorija Kukuškina, eQTLGen Consortium, Kai Kisand, Zoltán Kutalik, Pärt Peterson, and Hedi Peterson. C-reactive protein upregulates the whole blood expression of CD59 - an integrative analysis. *PLoS Comput. Biol.*, 13(9):e1005766, September 2017.

[2] Katrin Männik, Thomas Arbogast, Maarja Lepamets, Kaido Lepik, Anna Pellaz, Herta Ademi, Zachary A Kupchinsky, Jacob Ellegood, Catia Attanasio, Andrea Messina, et al. Leveraging biobank-scale rare and common variant analyses to identify ASPHD1 as the main driver of reproductive traits in the 16p11.2 locus. *bioRxiv*, July 2019.

[3] Eleonora Porcu, Sina Rüeger, Kaido Lepik, eQTLGen Consortium, BIOS Consortium, Federico A Santoni, Alexandre Reymond, and Zoltán Kutalik. Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat. Commun.*, 10(1):3300, July 2019.

[4] Eleonora Porcu, Annique Claringbould, Kaido Lepik, BIOS Consortium, Tom G Richardson, Federico A Santoni, Lude Franke, Alexandre Reymond, and Zoltán Kutalik. The role of gene expression on human sexual dimorphism: too early to call. *bioRxiv*, page 2020.04.15.042986, April 2020.

[5] Eleonora Porcu, Jennifer Sjaarda, Kaido Lepik, Cristian Carmeli, Liza Darrous, Jonathan Sulc, Ninon Mounier, and Zoltán Kutalik. Causal inference methods to integrate omics and complex traits. *Cold Spring Harb. Perspect. Med.*, August 2020.

[6] Glen James, Sulev Reisberg, Kaido Lepik, Nicholas Galwey, Paul Avillach, Liis Kolberg, Reedik Mägi, Tõnu Esko, Myriam Alexander, Dawn Waterworth, et al. An exploratory phenome wide association study linking asthma and liver disease genetic variants to electronic health records from the Estonian Biobank. *PLoS One*, 14(4):e0215026, April 2019.

[7] Maarja Lepamets, Kaido Lepik, Tuuli Jürgenson, Mart Kals, C Carmeli, A Claringbould, M Bochud, S Stringhini, C Wijmenga, L Franke, et al. New CNV quality score enables discovering novel phenotype associations from genome-wide CNV analysis. *In preparation*.

[8] Eileen M Crimmins. Lifespan and healthspan: Past, present, and promise. *Gerontologist*, 55(6):901–911, December 2015.

[9] Teresa Niccoli and Linda Partridge. Ageing as a risk factor for disease. *Curr. Biol.*, 22(17):R741–52, September 2012.

[10] Jim Oeppen and James W Vaupel. Demography. broken limits to life expectancy. *Science*, 296(5570):1029–1031, May 2002.

[11] Linda Partridge, Joris Deelen, and P Eline Slagboom. Facing up to the global challenges of ageing. *Nature*, 561(7721):45–56, September 2018.

[12] Tinca J C Polderman, Beben Benyamin, Christiaan A de Leeuw, Patrick F Sullivan, Arjen van Bochoven, Peter M Visscher, and Danielle Posthuma. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.*, 47(7):702–709, July 2015.

[13] J C Venter, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, M Yandell, C A Evans, R A Holt, et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, February 2001.

[14] E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

[15] International HapMap Consortium. The international HapMap project. *Nature*, 426(6968):789–796, December 2003.

[16] 1000 Genomes Project Consortium, Gonçalo R Abecasis, David Altshuler, Adam Auton, Lisa D Brooks, Richard M Durbin, Richard A Gibbs, Matt E Hurles, and Gil A McVean. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, October 2010.

[17] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of GWAS discovery: Biology, function, and translation. *Am. J. Hum. Genet.*, 101(1):5–22, July 2017.

[18] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, et al. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). *Nucleic Acids Res.*, 45(D1):D896–D901, January 2017.

[19] Daniel J Schaid, Wenan Chen, and Nicholas B Larson. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.*, May 2018.

[20] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.*, 47(3):291–295, March 2015.

[21] Evan A Boyle, Yang I Li, and Jonathan K Pritchard. An expanded view of complex traits: From polygenic to omnigenic. *Cell*, 169(7):1177–1186, June 2017.

[22] Ellen M McDonagh, Michelle Whirl-Carrillo, Yael Garten, Russ B Altman, and Teri E Klein. From pharmacogenomic knowledge acquisition to clinical applications: the PharmGKB as a clinical pharmacogenomic biomarker resource. *Biomark. Med.*, 5(6):795–806, December 2011.

[23] Richard M Weinshilboum and Liewei Wang. Pharmacogenomics: Precision medicine and drug response. *Mayo Clin. Proc.*, 92(11):1711–1722, November 2017.

[24] Naomi R Wray, Michael E Goddard, and Peter M Visscher. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.*, 17(10):1520–1528, October 2007.

[25] Alicia R Martin, Christopher R Gignoux, Raymond K Walters, Genevieve L Wojcik, Benjamin M Neale, Simon Gravel, Mark J Daly, Carlos D Bustamante, and Eimear E Kenny. Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.*, 100(4):635–649, April 2017.

[26] Alicia R Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M Neale, and Mark J Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.*, 51(4):584–591, April 2019.

[27] Bingxin Zhao and Fei Zou. On prs for complex polygenic trait prediction. June 2019.

[28] Sulev Reisberg, Tatjana Iljasenko, Kristi Läll, Krista Fischer, and Jaak Vilo. Comparing distributions of polygenic risk scores of type 2 diabetes and coronary heart disease within different populations. *PLoS One*, 12(7):e0179238, July 2017.

[29] E S Lander and N J Schork. Genetic dissection of complex traits. *Science*, 265(5181):2037–2048, September 1994.

[30] George Davey Smith and Shah Ebrahim. 'mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.*, 32(1):1–22, 2003.

[31] Hugues Aschard, Bjarni J Vilhjálmsson, Amit D Joshi, Alkes L Price, and Peter Kraft. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *Am. J. Hum. Genet.*, 96(2):329–339, February 2015.

[32] Michael Chong, Jennifer Sjaarda, Marie Pigeyre, Pedrum Mohammadi-Shemirani, Ricky Lali, Ashkan Shoamanesh, Hertzel Chaim Gerstein, and Guillaume Paré. Novel drug targets for ischemic stroke identified through mendelian randomization analysis of the blood proteome. *Circulation*, June 2019.

[33] Andrew D Bretherick, Oriol Canela-Xandri, Peter K Joshi, David W Clark, Konrad Rawlik, Thibaud S Boutin, Yanni Zeng, Carmen Amador, Pau

Navarro, Igor Rudan, et al. Proteome-by-phenome mendelian randomisation detects 38 proteins with causal roles in human diseases and traits. May 2019.

[34] Andrew Plump and George Davey Smith. Identifying and validating new drug targets for stroke and beyond. *Circulation*, September 2019.

[35] David Moher, Sally Hopewell, Kenneth F Schulz, Victor Montori, Peter C Gøtzsche, P J Devereaux, Diana Elbourne, Matthias Egger, Douglas G Altman, and CONSORT. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *Int. J. Surg.*, 10(1):28–55, 2012.

[36] Nancy Cartwright. Are RCTs the gold standard? *Biosocieties*, 2(1):11–20, March 2007.

[37] George Davey Smith, Lavinia Paternoster, and Caroline Relton. When will mendelian randomization become relevant for clinical practice and public health? *JAMA*, 317(6):589–591, February 2017.

[38] Liis Leitsalu, Toomas Haller, Tõnu Esko, Mari-Liis Tammesoo, Helene Alavere, Harold Snieder, Markus Perola, Pauline C Ng, Reedik Mägi, Lili Milani, et al. Cohort profile: Estonian biobank of the estonian genome center, university of tartu. *Int. J. Epidemiol.*, 44(4):1137–1147, August 2015.

[39] Marylyn D Ritchie, Emily R Holzinger, Ruowang Li, Sarah A Pendergrass, and Dokyoon Kim. Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.*, 16(2):85–97, February 2015.

[40] Judea Pearl. Causal inference in statistics: An overview. *Stat. Surv.*, 3:96–146, 2009.

[41] Kyoko Watanabe, Sven Stringer, Oleksandr Frei, Maša Umićević Mirkov, Christiaan de Leeuw, Tinca J C Polderman, Sophie van der Sluis, Ole A Andreassen, Benjamin M Neale, and Danielle Posthuma. A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.*, August 2019.

[42] Dan L Nicolae, Eric Gamazon, Wei Zhang, Shiwei Duan, M Eileen Dolan, and Nancy J Cox. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.*, 6(4):e1000888, April 2010.

[43] Farhad Hormozdiari, Emrah Kostem, Eun Yong Kang, Bogdan Pasaniuc, and Eleazar Eskin. Identifying causal variants at loci with multiple signals of association. *Genetics*, 198(2):497–508, October 2014.

[44] Christian Benner, Chris C A Spencer, Aki S Havulinna, Veikko Salomaa, Samuli Ripatti, and Matti Pirinen. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, 32(10):1493–1501, May 2016.

[45] Mads Engel Hauberg, Wen Zhang, Claudia Giambartolomei, Oscar Franzén, David L Morris, Timothy J Vyse, Arno Ruusalepp, CommonMind Consortium, Pamela Sklar, Eric E Schadt, et al. Large-Scale identification of common trait and disease variants affecting gene expression. *Am. J. Hum. Genet.*, 100(6):885–894, June 2017.

[46] Michael D Gallagher and Alice S Chen-Plotkin. The Post-GWAS era: From association to function. *Am. J. Hum. Genet.*, 102(5):717–730, May 2018.

[47] Halit Ongen, Andrew A Brown, Olivier Delaneau, Nikolaos I Panousis, Alexandra C Nica, GTEx Consortium, and Emmanouil T Dermitzakis. Estimating the causal tissues for complex traits and diseases. *Nat. Genet.*, 49(12):1676–1683, December 2017.

[48] Claudia Giambartolomei, Damjan Vukcevic, Eric E Schadt, Lude Franke, Aroon D Hingorani, Chris Wallace, and Vincent Plagnol. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.*, 10(5):e1004383, May 2014.

[49] Farhad Hormozdiari, Martijn van de Bunt, Ayellet V Segrè, Xiao Li, Jong Wha J Joo, Michael Bilow, Jae Hoon Sul, Sriram Sankararaman, Bogdan Pasaniuc, and Eleazar Eskin. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.*, 99(6):1245–1260, December 2016.

[50] Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, GTEx Consortium, Dan L Nicolae, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.*, 47(9):1091–1098, September 2015.

[51] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda W J H Penninx, Rick Jansen, Eco J C de Geus, Dorret I Boomsma, Fred A Wright, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.*, 48(3):245–252, March 2016.

[52] Alvaro N Barbeira, Scott P Dickinson, Rodrigo Bonazzola, Jiamao Zheng, Heather E Wheeler, Jason M Torres, Eric S Torstenson, Kaanan P Shah, Tzintzuni Garcia, Todd L Edwards, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.*, 9(1):1825, May 2018.

[53] Nicholas Mancuso, Malika K Freund, Ruth Johnson, Huwenbo Shi, Gleb Kichaev, Alexander Gusev, and Bogdan Pasaniuc. Probabilistic fine-mapping of transcriptome-wide association studies. *Nat. Genet.*, 51(4):675–682, April 2019.

[54] Stephen Burgess and Simon G Thompson. Multivariable mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *Am. J. Epidemiol.*, 181(4):251–260, February 2015.

[55] Benjamin F Voight, Gina M Peloso, Marju Orho-Melander, Ruth Frikke-Schmidt, Maja Barbalic, Majken K Jensen, George Hindy, Hilma Hólm, Eric L Ding, Toby Johnson, et al. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *Lancet*, 380(9841):572–580, August 2012.

[56] Ron Do, Cristen J Willer, Ellen M Schmidt, Sebanti Sengupta, Chi Gao, Gina M Peloso, Stefan Gustafsson, Stavroula Kanoni, Andrea Ganna, Jin Chen, et al. Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nat. Genet.*, 45(11):1345–1352, November 2013.

[57] Bruce Alberts, Dennis Bray, Karen Hopkin, Alexander D Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Essential Cell Biology*. Garland Science, October 2013.

[58] Eva Bianconi, Allison Piovesan, Federica Facchin, Alina Beraudi, Raffaella Casadei, Flavia Frabetti, Lorenza Vitale, Maria Chiara Pelleri, Simone Tassani, Francesco Piva, et al. An estimation of the number of cells in the human body. *Ann. Hum. Biol.*, 40(6):463–471, November 2013.

[59] Steven L Salzberg. Open questions: How many genes do we have? *BMC Biol.*, 16(1):94, August 2018.

[60] Andrew L Hopkins and Colin R Groom. The druggable genome. *Nat. Rev. Drug Discov.*, 1(9):727–730, September 2002.

[61] Björn Schwanhäusser, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342, May 2011.

[62] Claudia Manzoni, Demis A Kia, Jana Vandrovcova, John Hardy, Nicholas W Wood, Patrick A Lewis, and Raffaele Ferrari. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Brief. Bioinform.*, 19(2):286–302, March 2018.

[63] 1000 Genomes Project Consortium, Adam Auton, Lisa D Brooks, Richard M Durbin, Erik P Garrison, Hyun Min Kang, Jan O Korbel, Jonathan L Marchini, Shane McCarthy, Gil A McVean, et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, October 2015.

[64] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, 38(8):904–909, August 2006.

[65] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, et al. Genes mirror geography within europe. *Nature*, 456(7218):98–101, November 2008.

[66] Matthew R Robinson, Aaron Kleinman, Mariaelisa Graff, Anna A E Vinkhuyzen, David Couper, Michael B Miller, Wouter J Peyrot, Abdel Abdellaoui, Brendan P Zietsch, Ilja M Nolte, et al. Genetic evidence of assortative mating in humans. *Nature Human Behaviour*, 1:0016, January 2017.

[67] Augustine Kong, Daniel F Gudbjartsson, Jesus Sainz, Gudrun M Jonsdottir, Sigurjon A Gudjonsson, Bjorgvin Richardsson, Sigrun Sigurdardottir, John Barnard, Bjorn Hallbeck, Gisli Masson, et al. A high-resolution recombination map of the human genome. *Nat. Genet.*, 31(3):241–247, July 2002.

[68] Anna L Tyler, Dana C Crawford, and Sarah A Pendergrass. The detection and characterization of pleiotropy: discovery, progress, and promise. *Brief. Bioinform.*, 17(1):13–22, January 2016.

[69] Urmo Võsa, Annique Claringbould, Harm-Jan Westra, Marc Jan Bonder, Patrick Deelen, Biao Zeng, Holger Kirsten, Ashis Saha, Roman Kreuzhuber, Silva Kasela, et al. Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. October 2018.

[70] Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.*, 11(7):459–463, July 2010.

[71] Gabriel E Hoffman. Correcting for population structure and kinship using the linear mixed model: theory and extensions. *PLoS One*, 8(10):e75707, October 2013.

[72] Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjálmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, et al. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.*, 47(3):284–290, March 2015.

[73] Wei Zhou, Jonas B Nielsen, Lars G Fritsche, Rounak Dey, Maiken E Gabrielsen, Brooke N Wolford, Jonathon LeFaive, Peter VandeHaar, Sarah A Gagliano, Aliya Gifford, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.*, August 2018.

[74] M Egger, G Davey Smith, M Schneider, and C Minder. Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109):629–634, September 1997.

[75] William G Cochran. The combination of estimates from different experiments. *Biometrics*, 10(1):101–129, 1954.

[76] R DerSimonian and N Laird. Meta-analysis in clinical trials. *Control. Clin. Trials*, 7(3):177–188, September 1986.

[77] International HapMap 3 Consortium, David M Altshuler, Richard A Gibbs, Leena Peltonen, David M Altshuler, Richard A Gibbs, Leena Peltonen, Emmanouil Dermitzakis, Stephen F Schaffner, Fuli Yu, et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, September 2010.

[78] Jie Huang, Bryan Howie, Shane McCarthy, Yasin Memari, Klaudia Walter, Josine L Min, Petr Danecek, Giovanni Malerba, Elisabetta Trabetti, Hou-Feng Zheng, et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.*, 6:8111, September 2015.

[79] Jouni Kuha and Colin Mills. On group comparisons with logistic regression models. *Sociol. Methods Res.*, January 2018.

[80] Luke R Lloyd-Jones, Matthew R Robinson, Jian Yang, and Peter M Visscher. Transformation of summary statistics from linear mixed model association on All-or-None traits to odds ratio. *Genetics*, 208(4):1397–1408, April 2018.

[81] Eric B Fauman, Praveen Surendran, Isobel D Stewart, Luca A Lotta, Karsten Suhre, Gabi Kastenmuller, John Danesh, Nicholas J Wareham, Adam Butterworth, and Claudia Langenberg. Using the history of biochemistry to illuminate the genetic architecture of metabolite GWAS and its implications for complex human phenotypes; (Abstract #294). Presented at the Annual Meeting of The American Society of Human Genetics, October 2019, Houston.

[82] Randall J Pruim, Ryan P Welch, Serena Sanna, Tanya M Teslovich, Peter S Chines, Terry P Gliedt, Michael Boehnke, Gonçalo R Abecasis, and Cristen J Willer. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*, 26(18):2336–2337, September 2010.

[83] Jian Yang, Teresa Ferreira, Andrew P Morris, Sarah E Medland, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Pamela A F Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.*, 44(4):369–75, S1–3, March 2012.

[84] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.*, 88(1):76–82, January 2011.

[85] Joseph K Pickrell. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.*, 94(4):559–573, April 2014.

[86] Gleb Kichaev and Bogdan Pasaniuc. Leveraging Functional-Annotation data in trans-ethnic Fine-Mapping studies. *Am. J. Hum. Genet.*, 97(2):260–271, August 2015.

[87] Christopher M Bishop. *Pattern recognition and machine learning*. Springer Science+ Business Media, 2006.

[88] Zhihong Zhu, Futao Zhang, Han Hu, Andrew Bakshi, Matthew R Robinson, Joseph E Powell, Grant W Montgomery, Michael E Goddard, Naomi R Wray, Peter M Visscher, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.*, 48(5):481–487, May 2016.

[89] Michael Wainberg, Nasa Sinnott-Armstrong, Nicholas Mancuso, Alvaro N Barbeira, David A Knowles, David Golan, Raili Ermel, Arno Ruusalepp, Thomas Quertermous, Ke Hao, et al. Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.*, 51(4):592–599, April 2019.

[90] Bogdan Pasaniuc, Noah Zaitlen, Huwenbo Shi, Gaurav Bhatia, Alexander Gusev, Joseph Pickrell, Joel Hirschhorn, David P Strachan, Nick Patterson, and Alkes L Price. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, 30(20):2906–2914, October 2014.

[91] Sina Rüeger, Aaron McDaid, and Zoltán Kutalik. Evaluation and application of summary statistic imputation to discover new height-associated loci. *PLoS Genet.*, 14(5):e1007371, May 2018.

[92] Tom G Richardson, Sean Harrison, Gibran Hemani, and George Davey Smith. An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. *Elife*, 8, March 2019.

[93] Lang Wu, Wei Shi, Jirong Long, Xingyi Guo, Kyriaki Michailidou, Jonathan Beesley, Manjeet K Bolla, Xiao-Ou Shu, Yingchang Lu, Qiuyin Cai, et al. A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nat. Genet.*, page 1, June 2018.

[94] Alexander Gusev, Nicholas Mancuso, Hyejung Won, Maria Kousi, Hilary K Finucane, Yakir Reshef, Lingyun Song, Alexias Safi, Steven McCarroll, Benjamin M Neale, et al. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nat. Genet.*, 50(4):538–548, April 2018.

[95] Amanda Dobbyn, Laura M Huckins, James Boocock, Laura G Sloofman, Benjamin S Glicksberg, Claudia Giambartolomei, Gabriel E Hoffman, Thanneer M Perumal, Kiran Girdhar, Yan Jiang, et al. Landscape of conditional eQTL in dorsolateral prefrontal cortex and co-localization with schizophrenia GWAS. *Am. J. Hum. Genet.*, 102(6):1169–1184, June 2018.

[96] Judea Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, 2000.

[97] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, December 1995.

[98] Judea Pearl. Comment: Graphical models, causality and intervention. *Statistical Science*, 8(3):266–269, 1993.

[99] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc., 1988.

[100] William H Greene. *Econometric analysis*. Pearson Prentice Hall, Upper Saddle River, NJ, 7th ed. edition, 2012.

[101] Jeffrey M Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2002.

[102] Jack Bowden, George Davey Smith, and Stephen Burgess. Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *Int. J. Epidemiol.*, 44(2):512–525, April 2015.

[103] Stephen Burgess and Simon G Thompson. Bias in causal estimates from mendelian randomization studies with weak instruments. *Stat. Med.*, 30(11):1312–1323, May 2011.

[104] John R Thompson, Cosetta Minelli, Jack Bowden, Fabiola M Del Greco, Dipender Gill, Elinor M Jones, Chin Yang Shapland, and Nuala A Sheehan. Mendelian randomization incorporating uncertainty about pleiotropy. *Statistics in Medicine*, 36(29):4627–4645, 2017.

[105] Jeffrey M Wooldridge. *Introductory Econometrics: A Modern Approach*. Cengage Learning, June 2013.

[106] Miguel A Hernán and James M Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, December 2020.

[107] Reuben M Baron and David A Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.*, 51(6):1173–1182, December 1986.

[108] Joshua Millstein, Bin Zhang, Jun Zhu, and Eric E Schadt. Disentangling molecular relationships with a causal inference test. *BMC Genet.*, 10:23, May 2009.

[109] Jason E Aten, Tova F Fuller, Aldons J Lusis, and Steve Horvath. Using genetic markers to orient the edges in quantitative trait networks: the NEO software. *BMC Syst. Biol.*, 2:34, April 2008.

[110] R C Richmond, G Hemani, K Tilling, G Davey Smith, and C L Relton. Challenges and novel approaches for investigating molecular mediation. *Hum. Mol. Genet.*, 25(R2):R149–R156, October 2016.

[111] Hyun Min Kang, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, and Eleazar Eskin. Efficient control of pop-

ulation structure in model organism association mapping. *Genetics*, 178(3):1709–1723, March 2008.

[112] George Thanassoulis and Christopher J O'Donnell. Mendelian randomization: Nature's randomized trial in the Post–Genome era. *JAMA*, 301(22):2386–2388, June 2009.

[113] Maurice J G Bun and Frank Windmeijer. A comparison of bias approximations for the 2SLS estimator. June 2011.

[114] Marie-Jo A Brion, Konstantin Shakhbazov, and Peter M Visscher. Calculating statistical power in mendelian randomization studies. *Int. J. Epidemiol.*, 42(5):1497–1501, October 2013.

[115] Stephen Burgess, Simon G Thompson, and CRP CHD Genetics Collaboration. Avoiding bias from weak instruments in mendelian randomization studies. *Int. J. Epidemiol.*, 40(3):755–764, June 2011.

[116] Stephen Burgess, Neil M Davies, and Simon G Thompson. Bias due to participant overlap in two-sample mendelian randomization. *Genet. Epidemiol.*, 40(7):597–608, 2016.

[117] Stephen Burgess, Adam Butterworth, and Simon G Thompson. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.*, 37(7):658–665, November 2013.

[118] GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, October 2017.

[119] C Bycroft, C Freeman, D Petkova, G Band, L T Elliott, and others. Genome-wide genetic data on 500,000 UK biobank participants. *bioRxiv*, 2017.

[120] Oriol Canela-Xandri, Konrad Rawlik, and Albert Tenesa. An atlas of genetic associations in UK biobank. *Nat. Genet.*, 50(11):1593–1599, November 2018.

[121] Stephen Burgess, Frank Dudbridge, and Simon G Thompson. Combining information on multiple instrumental variables in mendelian randomization: comparison of allele score and summarized data methods. *Stat. Med.*, 35(11):1880–1906, May 2016.

[122] Qin Qin Huang, Scott C Ritchie, Marta Brozynska, and Michael Inouye. Power, false discovery rate and winner's curse in eQTL studies. *Nucleic Acids Res.*, 46(22):e133, December 2018.

[123] Philip C Haycock, Stephen Burgess, Kaitlin H Wade, Jack Bowden, Caroline Relton, and George Davey Smith. Best (but oft-forgotten) practices:

the design, analysis, and interpretation of mendelian randomization studies. *Am. J. Clin. Nutr.*, 103(4):965–978, April 2016.

[124] Jack Bowden, Fabiola Del Greco M, Cosetta Minelli, Qingyuan Zhao, Debbie A Lawlor, Nuala A Sheehan, John Thompson, and George Davey Smith. Improving the accuracy of two-sample summary-data mendelian randomization: moving beyond the NOME assumption. *Int. J. Epidemiol.*, 48(3):728–742, June 2019.

[125] Stephen Burgess, Verena Zuber, Elsa Valdes-Marquez, Benjamin B Sun, and Jemma C Hopewell. Mendelian randomization with fine-mapped genetic data: choosing from large numbers of correlated instrumental variables. *arXiv*, July 2017.

[126] Zhihong Zhu, Zhili Zheng, Futao Zhang, Yang Wu, Maciej Trzaskowski, Robert Maier, Matthew R Robinson, John J McGrath, Peter M Visscher, Naomi R Wray, et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.*, 9(1):224, January 2018.

[127] Richard Barfield, Helian Feng, Alexander Gusev, Lang Wu, Wei Zheng, Bogdan Pasaniuc, and Peter Kraft. Transcriptome-wide association studies accounting for colocalization using egger regression. *Genet. Epidemiol.*, 42(5):418–433, July 2018.

[128] Stephen Burgess and Simon G Thompson. Use of allele scores as instrumental variables for mendelian randomization. *Int. J. Epidemiol.*, 42(4):1134–1144, August 2013.

[129] Gibran Hemani, Jack Bowden, and George Davey Smith. Evaluating the potential role of pleiotropy in mendelian randomization studies. *Hum. Mol. Genet.*, May 2018.

[130] Marie Verbanck, Chia-Yen Chen, Benjamin Neale, and Ron Do. Detection of widespread horizontal pleiotropy in causal relationships inferred from mendelian randomization between complex traits and diseases. *Nat. Genet.*, April 2018.

[131] Stephen Burgess and Simon G Thompson. Interpreting findings from mendelian randomization using the MR-Egger method. *Eur. J. Epidemiol.*, May 2017.

[132] Jack Bowden, George Davey Smith, Philip C Haycock, and Stephen Burgess. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.*, 40(4):304–314, May 2016.

[133] Jack Bowden, Fabiola Del Greco M, Cosetta Minelli, George Davey Smith, Nuala A Sheehan, and John R Thompson. Assessing the suitability of summary data for two-sample mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. *Int. J. Epidemiol.*, 45(6):1961–1974, December 2016.

[134] Fernando Pires Hartwig, George Davey Smith, and Jack Bowden. Robust inference in summary data mendelian randomization via the zero modal pleiotropy assumption. *Int. J. Epidemiol.*, July 2017.

[135] François Aguet, Alvaro N Barbeira, Rodrigo Bonazzola, Andrew Brown, Stephane E Castel, Brian Jo, Silva Kasela, Sarah Kim-Hellmuth, Yanyu Liang, Meritxell Oliva, et al. The GTEx consortium atlas of genetic regulatory effects across human tissues. *bioRxiv*, page 787903, October 2019.

[136] Aaron F McDaid, Peter K Joshi, Eleonora Porcu, Andrea Komljenovic, Hao Li, Vincenzo Sorrentino, Maria Litovchenko, Roel P J Bevers, Sina Rüeger, Alexandre Reymond, et al. Bayesian association scan reveals loci associated with human lifespan and linked biomarkers. *Nat. Commun.*, 8:15842, July 2017.

[137] Eleanor Sanderson, George Davey Smith, Frank Windmeijer, and Jack Bowden. An examination of multivariable mendelian randomization in the single-sample and two-sample summary data settings. *International Journal of Epidemiology*, 48(3):713–727, 2019.

[138] Steven Black, Irving Kushner, and David Samols. C-reactive protein. *J. Biol. Chem.*, 279(47):48487–48490, November 2004.

[139] C Franceschi, M Bonafè, S Valensin, F Olivieri, M De Luca, E Ottaviani, and G De Benedictis. Inflamm-aging. an evolutionary perspective on immunosenescence. *Ann. N. Y. Acad. Sci.*, 908:244–254, June 2000.

[140] P M Ridker, M J Stampfer, and N Rifai. Novel risk factors for systemic atherosclerosis: a comparison of c-reactive protein, fibrinogen, homocysteine, lipoprotein(a), and standard cholesterol screening as predictors of peripheral arterial disease. *JAMA*, 285(19):2481–2485, May 2001.

[141] A D Pradhan, J E Manson, N Rifai, J E Buring, and P M Ridker. C-reactive protein, interleukin 6, and risk of developing type 2 diabetes mellitus. *JAMA*, 286(3):327–334, July 2001.

[142] Christine M Albert, Jing Ma, Nader Rifai, Meir J Stampfer, and Paul M Ridker. Prospective study of c-reactive protein, homocysteine, and plasma lipid levels as predictors of sudden cardiac death. *Circulation*, 105(22):2595–2599, June 2002.

[143] Eric E Schadt, John Lamb, Xia Yang, Jun Zhu, Steve Edwards, Debraj Guhathakurta, Solveig K Sieberts, Stephanie Monks, Marc Reitman, Chunsheng Zhang, et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.*, 37(7):710–717, July 2005.

[144] Abbas Dehghan, Josée Dupuis, Maja Barbalic, Joshua C Bis, Gudny Eiriksdottir, Chen Lu, Niina Pellikka, Henri Wallaschofski, Johannes Kettunen, Peter Henneman, et al. Meta-analysis of genome-wide association studies in >80 000 subjects identifies multiple loci for c-reactive protein levels. *Circulation*, 123(7):731–738, February 2011.

[145] Rozenn N Lemaitre, Toshiko Tanaka, Weihong Tang, Ani Manichaikul, Millennia Foy, Edmond K Kabagambe, Jennifer A Nettleton, Irena B King, Lu-Chen Weng, Sayanti Bhattacharya, et al. Genetic loci associated with plasma phospholipid n-3 fatty acids: a meta-analysis of genome-wide association studies from the CHARGE consortium. *PLoS Genet.*, 7(7):e1002193, July 2011.

[146] Dariush Mozaffarian, Edmond K Kabagambe, Catherine O Johnson, Rozenn N Lemaitre, Ani Manichaikul, Qi Sun, Millennia Foy, Lu Wang, Howard Wiener, Marguerite R Irvin, et al. Genetic loci associated with circulating phospholipid trans fatty acids: a meta-analysis of genome-wide association studies from the CHARGE consortium. *Am. J. Clin. Nutr.*, 101(2):398–406, February 2015.

[147] Nicola Martinelli, Domenico Girelli, Giovanni Malerba, Patrizia Guarini, Thomas Illig, Elisabetta Trabetti, Marco Sandri, Simonetta Friso, Francesca Pizzolo, Linda Schaeffer, et al. FADS genotypes and desaturase activity estimated by the ratio of arachidonic acid to linoleic acid are associated with inflammation and coronary artery disease. *Am. J. Clin. Nutr.*, 88(4):941–949, October 2008.

[148] A Davies, D L Simmons, G Hale, R A Harrison, H Tighe, P J Lachmann, and H Waldmann. CD59, an LY-6-like protein expressed in human lymphoid cells, regulates the action of the complement membrane attack complex on homologous cells. *J. Exp. Med.*, 170(3):637–654, September 1989.

[149] Harm-Jan Westra, Marjolein J Peters, Tõnu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W Christiansen, Benjamin P Fairfax, Katharina Schramm, Joseph E Powell, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.*, 45(10):1238–1243, October 2013.

[150] Sébastien Jacquemont, Alexandre Reymond, Flore Zufferey, Louise Harewood, Robin G Walters, Zoltán Kutalik, Danielle Martinet, Yiping Shen, Armand Valsesia, Noam D Beckmann, et al. Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. *Nature*, 478(7367):97–102, August 2011.

[151] Lauren A Weiss, Yiping Shen, Joshua M Korn, Dan E Arking, David T Miller, Ragnheidur Fossdal, Evald Saemundsen, Hreinn Stefansson, Manuel A R Ferreira, Todd Green, et al. Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.*, 358(7):667–675, February 2008.

[152] A M Maillard, A Ruef, F Pizzagalli, E Migliavacca, L Hippolyte, S Adaszewski, J Dukart, C Ferrari, P Conus, K Männik, et al. The 16p11.2 locus modulates brain structures common to autism, schizophrenia and obesity. *Mol. Psychiatry*, 20(1):140–147, February 2015.

[153] Loyse Hippolyte, Anne M Maillard, Borja Rodriguez-Herreros, Aurélie Pain, Sandra Martin-Brevet, Carina Ferrari, Philippe Conus, Aurélien Macé, Nouchine Hadjikhani, Andres Metspalu, et al. The number of genomic copies at the 16p11.2 locus modulates language, verbal memory, and inhibition. *Biological Psychiatry*, 80(2):129–139, 2016.

[154] Felix R Day, Deborah J Thompson, Hannes Helgason, Daniel I Chasman, Hilary Finucane, Patrick Sulem, Katherine S Ruth, Sean Whalen, Abhishek K Sarkar, Eva Albrecht, et al. Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. *Nat. Genet.*, 49(6):834–841, June 2017.

[155] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, May 2020.

[156] Ting Qi, Yang Wu, Jian Zeng, Futao Zhang, Angli Xue, Longda Jiang, Zhihong Zhu, Kathryn Kemper, Loic Yengo, Zhili Zheng, et al. Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat. Commun.*, 9(1):2282, June 2018.

[157] E Abraham, O Palevitch, S Ijiri, S J Du, Y Gothilf, and Y Zohar. Early development of forebrain gonadotrophin-releasing hormone (GnRH) neurones and the role of GnRH as an autocrine migration factor. *J. Neuroendocrinol.*, 20(3):394–405, 2008.

[158] Eytan Abraham, Ori Palevitch, Yoav Gothilf, and Yonathan Zohar. The zebrafish as a model system for forebrain GnRH neuronal development. *Gen. Comp. Endocrinol.*, 164(2-3):151–160, November 2009.

[159] Ori Palevitch, Katherine Kight, Eytan Abraham, Susan Wray, Yonathan Zohar, and Yoav Gothilf. Ontogeny of the GnRH systems in zebrafish brain: in situ hybridization and promoter-reporter expression analyses in intact animals. *Cell Tissue Res.*, 327(2):313–322, February 2007.

[160] Patrick D Hsu, Eric S Lander, and Feng Zhang. Development and applications of CRISPR-Cas9 for genome engineering. *Cell*, 157(6):1262–1278, June 2014.

[161] Nurlan Kerimov, James D Hayhurst, Jonathan R Manning, Peter Walter, Liis Kolberg, Kateryna Peikova, Marija Samoviča, Tony Burdett, Simon Jupp, Helen Parkinson, et al. eQTL catalogue: a compendium of uniformly processed human gene expression and splicing QTLs. January 2020.

[162] Encarna Guillén-Navarro, Sofía Sánchez-Iglesias, Rosario Domingo-Jiménez, Berta Victoria, Alejandro Ruiz-Riquelme, Alberto Rábano, Lourdes Loidi, Andrés Beiras, Blanca González-Méndez, Adriana Ramos, et al. A new seipin-associated neurodegenerative syndrome. *J. Med. Genet.*, 50(6):401–409, June 2013.

[163] Ranad Shaheen, Eissa Faqeih, Shinu Ansari, Ghada Abdel-Salam, Zuhair N Al-Hassnan, Tarfa Al-Shidi, Rana Alomar, Sameera Sogaty, and Fowzan S Alkuraya. Genomic analysis of primordial dwarfism reveals novel disease genes. *Genome Res.*, 24(2):291–299, February 2014.

[164] Andrew R Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan, Zoltán Kutalik, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.*, 46(11):1173–1186, November 2014.

[165] Melina Claussnitzer, Simon N Dankel, Kyoung-Han Kim, Gerald Quon, Wouter Meuleman, Christine Haugen, Viktoria Glunk, Isabel S Sousa, Jacqueline L Beaudry, Vijitha Puviindran, et al. FTO obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.*, 373(10):895–907, September 2015.

[166] Kiran Musunuru, Alanna Strong, Maria Frank-Kamenetsky, Noemi E Lee, Tim Ahfeldt, Katherine V Sachs, Xiaoyu Li, Hui Li, Nicolas Kuperwasser, Vera M Ruda, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*, 466(7307):714–719, August 2010.

[167] Ekaterina A Khramtsova, Lea K Davis, and Barbara E Stranger. The role of sex in the genomics of human complex traits. *Nat. Rev. Genet.*, December 2018.

[168] Alison M Kim, Candace M Tingen, and Teresa K Woodruff. Sex bias in trials and treatment must end. *Nature*, 465(7299):688–689, June 2010.

[169] Michael E Sobel. Asymptotic confidence intervals for indirect effects in structural equation models. *Sociol. Methodol.*, 13:290–312, 1982.

[170] Jie Zheng, A Mesut Erzurumluoglu, Benjamin L Elsworth, John P Kemp, Laurence Howe, Philip C Haycock, Gibran Hemani, Katherine Tansey, Charles Laurin, Beate St. Pourcain, et al. LD hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis, 2017.

[171] Mihir A Kamat, James A Blackshaw, Robin Young, Praveen Surendran, Stephen Burgess, John Danesh, Adam S Butterworth, and James R Staley. PhenoScanner v2: an expanded tool for searching human genotype–phenotype associations. *Bioinformatics*, 35(22):4851–4853, November 2019.

[172] Gibran Hemani, Jie Zheng, Benjamin Elsworth, Kaitlin H Wade, Valeriia Haberland, Denis Baird, Charles Laurin, Stephen Burgess, Jack Bowden, Ryan Langdon, et al. The MR-Base platform supports systematic causal inference across the human phenome. *Elife*, 7, May 2018.

[173] Michael V Holmes, Mika Ala-Korpela, and George Davey Smith. Mendelian randomization in cardiometabolic disease: challenges in evaluating causality. *Nat. Rev. Cardiol.*, June 2017.

[174] Louise A C Millard, Neil M Davies, Kate Tilling, Tom R Gaunt, and George Davey Smith. Searching for the causal effects of body mass index in over 300 000 participants in UK biobank, using mendelian randomization, 2019.

[175] Jie Zheng, Valeriia Haberland, Denis Baird, Venexia Walker, Philip C Haycock, Mark R Hurle, Alex Gutteridge, Pau Erola, Yi Liu, Shan Luo, et al. Phenome-wide mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat. Genet.*, 52(10):1122–1131, October 2020.

[176] Brielin C Brown and David A Knowles. Phenome-scale causal network discovery with bidirectional mediated mendelian randomization. June 2020.

ACKNOWLEDGEMENT

> It is better to deserve honors and not have them than to have them and not to deserve them
>
> Mark Twain

What was difficult and frustrating to cope with in the beginning felt at times almost like a privilege by the end. It would not have been possible without financial support. For that, I am thankful to Hedi, Jaak and Lili. I am most grateful to Hedi for being available, sharing her experience, being proactive in dealing with whatever needed to be dealt with, and for keeping me in check at times. I am also indebted to Pärt for his valuable mentorship early in my PhD.

I would like to thank the many people who proofread this dissertation and suggested useful comments. This includes my reviewer Kaur and oponents Prof. Jack Bowden and Prof. Samuli Ripatti whose excellent ideas and suggestions helped me to piece everything together. My special gratitude goes to Jüri who selflessly went the extra mile to help me refine mathematical clarity—the rigour of his advice was a privilege to experience and benefit from.

Scientifically, I learned a lot from Zoltán and am indebted to him for having me in his Statistical Genetics Group in Lausanne. The five months spent in Lausanne and on many hiking trails all over Switzerland in the weekends represents arguably the best time of my life. I grew a lot in Switzerland and gained confidence to pursue my PhD in causal inference. Zoltán's influence cannot be overstated here. It is also hard to describe how grateful I am to have met so many wonderful people in his group—Sina, Jonathan, Ninon, Aaron, Anthony, Eleonora—who welcomed me with open arms, introduced me to the local life, and helped me learn the ways of statistical genetics. I will always be thankful to Katrin who—at the time not having even met me—helped me get there by lending a good word to the Swiss people on my behalf, supported me in writing the scholarship application, and shared her experiences about life in Switzerland. I am happy to have met Ott who, with no benefit to himself, took the trouble to introduce me to downhill skiing and show the best places to visit in Geneva.

Daily, I have shared my PhD experience with the wonderful people of the BIIT group. Some of them have graduated before me and shown me how it is done; some are still walking the path. I am grateful for all the fun conversations we had over banana time and I want to thank you all! Elena for her positivity and enthusiasm; Priit (Lemps) for being grounded and wise; Dima for his endless insights and initiatives (and table tennis practices!); Liis for her ethics and staying true to beliefs; Ivan for his genius and incredible kindness towards others; Uku for his constructive and rational reasoning; Erik for his experience and knowledge about the world; Mari-Liis for her approachability and can-do attitude; Nurlan for his down-to-earth ambition and sincere/focused demeanour; and others too

plentiful to name here. Alas, I am most thankful to Kaur, Ahto, and Kateryna.

It is hard to imagine a more wholesome person than Kaur and I will not tire to say it to anyone willing to listen. :) Thank you for the discussions, the ideas, the positive attitude, accessibility and curiosity, finding time for others, sharing knowledge, and leading by example. Your impact on the young'uns in science has been tremendous and gives me faith for great things to come. Ahto's appetite for knowledge, ability to store information and to apply logical reasoning down to fundamental principles is second to none. I am thankful for the insightful conversations, the biological know-how, the brutal honesty, and of course the gym sessions. Towards the end of my PhD, I am happy to have met Kate. Always smiling and energetic, every day lifting the mood around the office. The avocado, the guitar, all the other crazy stuff.. I could not thank you enough, Katya!

During my PhD, I was also associated with the Estonian Genome Center. I admire Andres' inspiring leadership in developing the genome program and his appreciation towards the people who contribute to make it happen; Reedik's and Mart's lasting willingness to provide answers to anyone searching for the light; Krista for propagating solid statistics and science; and all the others with whom I had the pleasure to connect with. I am most thankful to Urmo for selflessly making my Groningen visit well organised and worthwhile.

I have only ever supervised one student but the bar was raised so incredibly high that I could probably never supervise anyone ever again. :) Thank you for this great collaboration, Tuuli! It is difficult to say who actually learned more from this; thank you for restoring my faith in always putting an effort in and striving to do better in whatever task at hand!

There could not have been a healthy PhD experience without ways to escape from work-related thoughts. For me, this came down to practicing sports. Thank you, coaches Andres, Rahel and Jaak, for the chill and relaxed environment in training—I was always looking forward to this much-needed change in atmoshpere. Thank you, fellow Suusahullud: Laura for being so active and pulling others along; Sille and Janno for the fighting spirit in skiing, volleyball and Xdream alike; Mart for the marathon prep. Likewise, thank you Kersti, Madis and Lauri for all the sports—be it be running, orienteering, or rogaining—and the excited discussions that inevitably both preceded and followed every competition.

I am also grateful to my family for being there for me when it mattered the most. Thank you, Dad, for letting me choose my own path in life while providing a place to fall back on when necessary. Likewise, thank you, Gea, for being in many ways the most extraordinary person I know. If ever there was someone capable of achieving anything they set their mind on, it is my sister.

For two years of my studies, Käbi was part of my everyday life, providing happiness and joy. The amount of positive energy she emanated, specifically every morning after waking up and every evening after coming home, was instrumental. I miss it. We had so much fun as we trained, ran, hiked, played, wrestled, and altogether grew up together. You will forever be in my heart.

SISUKOKKUVÕTE

Haiguspõhjuslike geenide tuvastamine statistiliste meetoditega

Üheks suurimaks inimkonna saavutuseks viimasel sajandil on olnud keskmise eluea kiire kasv. Tõusnud on ka keskmine tervena elatud aastate arv, ent mitte võrdväärses tempos. Probleemiks on just vanusega kaasnevad kroonilised haigused (nt südame- ja veresoonkonnahaigused). Tervislikke eluviise järgides on võimalik haigusi ennetada, kuid pea sama oluline roll on ka geneetilistel protsessidel. Tänapäeva meditsiin ekspluateerib seda teadmist, arendades ravimeid, mis korraldavad haigusega seotud geeniproduktide ehk valkude töö haigusele pärssivalt ümber. Selleks on esmalt tarvis leida haiguspõhjuslikud geenid, mille produktide funktsiooni modifitseerida. Peamiseks standardiks ravimite tööpõhimõtte valideerimiseks (tegelikult igasuguse põhjusliku seose uurimiseks) on kontrollgrupiga kliinilised uuringud. Kuigi sellised uuringud on ravimitööstuses laialt levinud, on kogu protsessi läbiviimine üsna kulukas ja aeganõudev. Liiati ei ole kliiniliste uuringute tegemine eetilistel kaalutlustel alati võimalikki. Protsessi efektiivsust on võimalik oluliselt tõsta, kui prioritiseerida uuringutes vaadeldavaid kandidaatgeene näiteks statistiliste meetoditega.

Käesolevas doktoritöös otsimegi statistilise analüüsi abil haigusi ja teisi kompleksseid fenotüübilisi tunnuseid põhjuslikult mõjutavaid geene. Tegu on kiiresti areneva teadusvaldkonnaga statistilises geneetikas, mis on hoo sisse saanud tänu rahvuslike biopankade tegevusele, mille tulemusel on tekkinud suured andmehulgad inimeste geno- ja fenotüüpidega (nt Tartu Ülikooli Eesti geenivaramuga on liitunud ligi viiendik eestlastest). Samas on matemaatiline raamistik põhjuslike seoste uurimiseks alles arenemisjärgus. Analüüsiks vajalik teooria hõlmab laene erinevatelt teadusaladelt – traditsioonilisest statistikast, ökonomeetriast, geneetikast, põhjuslikkuse teooriast –, millel puudub ühtne käsitlus. Selle doktoritöö üheks põhieesmärgiks ja panuseks (lisaks publitseeritud teadusartiklites loodud uuele teadmisele) on vastav teooria haiguspõhjuslike geenide leidmise kontekstis harmoniseerida. Eesmärgi realiseerimiseks pühendame märkimisväärselt palju aega matemaatilise teooria põhjalikule käsitlemisele, tutvustades olulisi kontseptsioone, statistilisi meetodeid ja oskusteavet. Alles seejärel kirjeldame doktoritöö raames valminud teaduspublikatsioonide tulemusi, mis põhinevad eelnevalt tutvustatud metoodika arendamisel ja rakendamisel praktikas inimeste andmetel.

Paljude fenotüübiliste tunnuste (sh üldlevinud haiguste) väljakujunemine on kompleksne protsess. Rolli võivad omada paljud erinevad geenid üle kogu genoomi, kusjuures igal üksikul geenil võib olla vaid marginaalne mõju. Väikeste põhjuslike efektide avastamiseks on üldjuhul vajalikud suured, tuhandetesse ulatuvad valimimahud. Oma teadustöö raames arendasime välja metoodika, mis erinevate põhjuslike seoste hulgast suurima tõepära printsiibil kõige usaldusväärsemat seost valides võimaldab funktsionaalseid geene prioritiseerida ka väiksema valimimahu

korral ($n \approx 500$). Rakendasime antud metoodikat Eesti geenivaramu andmetel, et uurida põletikumarkeri C-reaktiivse valgu funktsiooni põletikuprotsessides. Leidsime, et see reguleerib geeni *CD59* avaldumist veres. Kuna vastav geen inhibeerib immuunreaktsiooni tugevust, võib C-reaktiivne valk osaleda tervete vererakkude kaitsmisel organismi immuunvastuses patogeenidele.

Iga statistiline meetod põhineb teatud eeldustele uuritava fenomeni kohta, mille paikapidavusest sõltub analüüsi järelduste usaldusväärsus. Põhjuslike geenide leidmise kontekstis on väga oluline arvestada DNA pärilikkusseadustega. Näiteks kipuvad lähedalasuvad genoomipiirkonnad päranduma koos, mis tekitab nendevahelist seost. Probleemiks on ka laialdaselt levinud pleiotroopia – nähtus, mille kohaselt võib üks geen mõjutada paljusid erinevaid tunnuseid. Mõlemad seaduspärad raskendavad põhjuslikku interpretatsiooni. Teadustöö raames töötasime välja nende nähtuste osas robustse (võrreldes alternatiividega) algoritmi, mis võimaldab mistahes tunnuse kujunemist mõjutavaid geene leida üle kogu genoomi. Suurema haarde saavutamiseks geen-tunnus vaheliste seoste kirjeldamiseks rakendasime seda Mendeli pärilikkusseadustel põhinevat metoodikat 43 erineval fenotüübilisel tunnusel. Leidsime tuhandeid uusi seoseid. Seejuures näitasime väljatöötatud metoodika ülimuslikkust võrreldes alternatiividega, seda nii suurema statistilise võimsuse kui ka väiksema I tüüpi vea tegemise protsendi osas. Eriti põhjalikult uurisime antud metoodikaga ühe iseäranis probleemse ja geenitiheda genoomipiirkonna (16p11.2) seost inimeste seksuaalse arenguga. Viimane on seotud haigustega hilisemas elus, seega on vastavate protsesside mõistmine olulise tähtsusega. Analüüsi tulemusel suutsime osutada geenidele (*ASPHD1*, *KCTD13*), mis omavad põhjuslikku mõju.

Meditsiinis on järjest enam kandepinda võtmas personaalsed, iga inimese (geneetiliste) eripäradega arvestavad lahendused. Erilist tähelepanu on pälvinud ravimite ja ravimiannuste määramine vastavalt inimeste ainevahetuslikele iseärasustele. Samas ilmnevad erinevused fenotüübilistes tunnustes juba ühiskonnakihtide lõikes, näiteks meeste ja naiste vahel, ja personaalmeditsiini juurutamisele võib nende erinevustega arvestamine tähendada tõelist läbimurret. Seetõttu uurisime doktoritöös käsitletavas teadustöös ka seda, mil määral on fenotüübiliste tunnuste soospetsiifilisus tingitud eripäradest geenide avaldumises ning kas meestel ja naistel võivad rolli mängida erinevad põhjuslikud geenid. Näitame statistilisele võimsusanalüüsile tuginedes, et arvutuslike meetoditega ei ole hetkel võimalik nendele küsimustele lõplikku hinnangut anda – selleks oleks tarvis suuremaid andmemahte, kui avalikus ruumis parasjagu kättesaadav on.

Viimaks, mistahes kahe tunnuse vahelise põhjusliku seose avastamiseks ei piisa üldjuhul korrelatsiooni leidmisest nende tunnuste vahel. Kuigi mõnel juhul võib see tõesti viidata funktsionaalsele seosele, võib korrelatsiooni tekkimine olla tingitud segavatest faktoritest – kolmandatest tunnustest, mis on seotud nii ühe kui ka teise uuritava tunnuse kujunemisega. Vastupidine siiski kehtib – põhjuslik lineaarne seos (üldjuhul me eeldamegi lineaarseid seoseid) tekitab tunnuste vahel ka korrelatsiooni. Huvitaval kombel ei näinud me oma teadustöös aga peaaegu

mitte mingisugust ülekatet geenide vahel, millel leiti vaatlusandmetel statistilisi meetodeid rakendades uuritavate tunnustega põhjuslik side või pelgalt korrelatsioon. Antud fenomen võib olla tingitud geenide omavahelisest tugevast struktuurist, mis läbi paljude segavate faktorite moonutab igasuguse põhjusliku signaali. Samas teoretiseerime oma teadustöös ka selle üle, et tunnustega korreleeritud geenidel võib olla parem ülekate nende geenidega, mille avaldumisele mõjuvad põhjuslikult hoopis komplekstunnused ise (nt haigused). Igal juhul näitab meie tähelepanek selgelt, et korrelatsioone ei ole mõistlik kasutada funktsionaalsete geenide prioritiseerimiseks.

Fikseeritud vaatlusandmete pealt ei ole üldjuhul võimalik kontrollida kõikide eelduste kehtivust, millele põhjuslikud mudelid tuginevad. See eeldaks eelteadmisi tunnustevahelise põhjusliku struktuuri kohta, mida teades ei oleks enam vajadust analüüsi teostadagi. Võimetus eelduste täitmist verifitseerida võib tekitada küsitavusi ka tulemuste ja järelduste paikapidavuse kohta. Doktoritöö raames publitseeritud teadustöödes oleme seda riski minimiseerinud, rakendades sama nähtuse uurimiseks erinevaid metoodikaid, sh valideerinud leitud tulemusi katseliselt laboris. Tunnustevahelisi seoseid oleme põhjuslikult interpreteerinud vaid siis, kui oleme erinevate lähenemistega jõudnud samadele järeldustele. Lõppsõna põhjuslike seoste kehtivuse osas jääb alati laborikatsetele või kontrollgrupiga kliinilistele uuringutele. Siiski on selge, et statistilised meetodid võimaldavad analüüsida suurt hulka andmeid väga kiiresti. Pakutav efektiivsus on võtmetähtsusega, et ravimiarenduse teel anda märkimisväärne panus haigustega võitlemisse.

PUBLICATIONS

CURRICULUM VITAE

Personal data

Name:	Kaido Lepik
Date of birth:	05.11.1990
Citizenship:	Estonia
E-mail:	kaido.lepik@ut.ee

Education

2015–...	University of Tartu, Faculty of Science and Technology, Computer Science, PhD
2012–2014	University of Tartu, Faculty of Mathematics and Computer Science, Mathematical Statistics, MSc (*cum laude*)
2009–2012	University of Tartu, Faculty of Mathematics and Computer Science, Mathematical Statistics, BSc (*cum laude*)

Employment

2018–...	University of Tartu, Institute of Genomics (Genome Center), Specialist for Bioinformatics
2018–...	University of Tartu, Institute of Computer Science, Junior Research Fellow in Bioinformatics
2017	Centre hospitalier universitaire vaudois in Lausanne, Switzerland, Statistical Genetics Group, Visiting PhD Student
Nov 2017	Universitair Medisch Centrum Groningen in the Netherlands, Functional Genomics Group, Visiting PhD Student
2012–2015	Kantar Emor, Data Engineer / Data Scientist

Honours & awards

Dec 2019	Supervised Tuuli Jürgenson to 2nd prize in the annual Estonian research National Contest for University Students with her BSc thesis "Associations between copy number variations and adverse drug reactions"
Nov 2019	1st place in the Estonian Bioinnovation Days 2019 hackaton with a DNA methylation based health monitoring tool
2017	Kristjan Jaak Scholarship for Short Study Visit for the research visit to Groningen, Netherlands
2016	Dora Plus PhD Student Mobility Scholarship for the research visit to Lausanne, Switzerland
2014	2nd place in the Student Project Contest of the Institute of Computer Science of the University of Tartu with the MSc project

Teaching

2019 University of Tartu, Faculty of Science and Technology, Institute of Computer Science, Teaching Assistant in Machine Learning 2

Supervised theses

2019 Tuuli Jürgenson, BSc thesis "Associations between copy number variations and adverse drug reactions"

Scientific work

Main fields of interest:

- statistical genetics – developing and applying statistical methods to learn about human genetics;
- omics integration – combining multiple layers of biological data to gain more insight into biological processes underlying complex trait variation;
- causal inference – studying and using analysis tools such as Mendelian randomization to unravel functional relationships and direction of effects between traits.

Publications and preprints

2020 Eleonora Porcu, Jennifer Sjaarda, **Kaido Lepik**, Cristian Carmeli, Liza Darrous, Jonathan Sulc, Ninon Mounier, and Zoltán Kutalik. Causal inference methods to integrate omics and complex traits. *Cold Spring Harb. Perspect. Med.*, August 2020.

2020 Eleonora Porcu, Annique Claringbould, **Kaido Lepik**, BIOS Consortium, Tom G Richardson, Federico A Santoni, Lude Franke, Alexandre Reymond, and Zoltán Kutalik. The role of gene expression on human sexual dimorphism: too early to call. *bioRxiv*, April 2020.

2019 Katrin Männik, Thomas Arbogast, Maarja Lepamets, **Kaido Lepik**, Anna Pellaz, Herta Ademi, Zachary A Kupchinsky, Jacob Ellegood, Catia Attanasio, Andrea Messina, Samuel Rotman, Sandra Martin-Brevet, Estelle Dubruc, Jacqueline Chrast, Jason P Lerch, Lily R Qiu, Triin Laisk, The 16p11.2 European Consortium, The Simons VIP Consortium, The eQTLGen Consortium, R Mark Henkelman, Sébastien Jacquemont, Yann Herault, Cecilia M Lindgren, Hedi Peterson, Jean Christophe Stehle, Nicholas Katsanis, Zoltan Kutalik, Serge Nef, Bogdan Draganski, Erica E Davis, Reedik Mägi, and Alexandre Reymond. Leveraging biobank-scale rare and common variant analyses to identify ASPHD1 as the main driver of reproductive traits in the 16p11.2 locus. *bioRxiv*, July 2019.

2019 Eleonora Porcu, Sina Rüeger, **Kaido Lepik**, eQTLGen Consortium, BIOS Consortium, Federico A Santoni, Alexandre Reymond, and Zoltán Kutalik. Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat. Commun.*, 10(1):3300, July 2019.

2019 Glen James, Sulev Reisberg, **Kaido Lepik**, Nicholas Galwey, Paul Avillach, Liis Kolberg, Reedik Mägi, Tõnu Esko, Myriam Alexander, Dawn Waterworth, A Katrina Loomis, and Jaak Vilo. An exploratory phenome wide association study linking asthma and liver diseasegenetic variants to electronic health records from the Estonian Biobank. *PLoS One*, 14(4):e0215026, April 2019.

2017 **Kaido Lepik**, Tarmo Annilo, Viktorija Kukuškina, eQTLGen Consortium, Kai Kisand, Zoltán Kutalik, Pärt Peterson, and Hedi Peterson. C-reactive protein upregulates the whole blood expression of CD59 - an integrative analysis. *PLoS Comput. Biol.*, 13(9):e1005766, September 2017.

TBD Maarja Lepamets, **Kaido Lepik**, Tuuli Jürgenson, Mart Kals, Cristian Carmeli, Annique Claringbould, Murielle Bochud, Silvia Stringhini, Cisca Wijmenga, Lude Franke, Reedik Mägi, and Zoltán Kutalik. New CNV quality score enables discovering novel phenotype associations from genome-wide CNV analysis. *In preparation*.

ELULOOKIRJELDUS

Isikuandmed

Nimi:	Kaido Lepik
Sünniaeg:	05.11.1990
Kodakondsus:	Eesti
E-mail:	kaido.lepik@ut.ee

Haridus

2015–...	Tartu Ülikool, loodus- ja täppisteaduste valdkond, informaatika, doktoriõpe
2012–2014	Tartu Ülikool, matemaatika-informaatikateaduskond, matemaatiline statistika, magistriõpe (*cum laude*)
2009–2012	Tartu Ülikool, matemaatika-informaatikateaduskond, matemaatiline statistika, bakalaureuseõpe (*cum laude*)

Teenistuskäik

2018–...	Tartu Ülikool, loodus- ja täppisteaduste valdkond, arvutiteaduse instituut, bioinformaatika nooremteadur
2018–2019	Tartu Ülikool, Tartu Ülikooli genoomika instituut (geenivaramu), bioinformaatika spetsialist
2017	Centre hospitalier universitaire vaudois Lausanne'is Šveitsis, statistilise geneetika grupp, külalisdoktorant
Nov 2017	Universitair Medisch Centrum Groningen Hollandis, funktsionaalse genoomika grupp, külalisdoktorant
2012–2015	Kantar Emor, andmeinsener / andmeteadlane

Teaduspreemiad ja tunnustused

Dets 2019	Juhendatava Tuuli Jürgensoni II preemia üliõpilaste teadustööde riiklikul konkursil bakalaureusetööga "Koopiaarvu variatsioonide mõju ravimi kõrvaltoimete tekkimisele"
Nov 2019	I koht Eesti Bioinnovatsiooni päevad 2019 häkatonil rakendusega inimeste tervisenäitajate jälgimiseks DNA metülatsiooni põhjal
2017	Kristjan Jaagu välislähetuste stipendium teaduskoostööks Groningenis Hollandis
2016	Dora Pluss T1.2 doktorantide õpirände stipendium teaduskoostööks Lausanne'is Šveitsis
2014	II koht Tartu Ülikooli arvutiteaduse instituudi tudengiprojektide võistlusel magistritööga vastavas kategoorias

Õppetöö

2019 Tartu Ülikool, loodus- ja täppisteaduste valdkond, arvutiteaduse instituut, õppeassistent aines masinõpe 2

Juhendatud väitekirjad

2019 Tuuli Jürgenson, bakalaureusetöö "Koopiaarvu variatsioonide mõju ravimi kõrvaltoimete tekkimisele"

Teadustegevus

Peamised uurimisvaldkonnad:

- statistiline geneetika – statistiliste meetodite arendamine ja kasutamine järelduste tegemiseks inimgeneetikas;
- oomikate integreerimine – erinevate bioloogiliste andmekihtide kombineerimine, saamaks täpsemat ja mitmekülgsemat infot komplekstunnuste varieeruvust põhjustavatest bioloogilistest protsessidest;
- põhjuslik analüüs – meetodite nagu Mendeli randomiseerimine arendamine ja rakendamine tunnustevaheliste põhjuslike seoste avastamiseks.

Publikatsioonid ja eeltrükid

2020 Eleonora Porcu, Jennifer Sjaarda, **Kaido Lepik**, Cristian Carmeli, Liza Darrous, Jonathan Sulc, Ninon Mounier, and Zoltán Kutalik. Causal inference methods to integrate omics and complex traits. *Cold Spring Harb. Perspect. Med.*, August 2020.

2020 Eleonora Porcu, Annique Claringbould, **Kaido Lepik**, BIOS Consortium, Tom G Richardson, Federico A Santoni, Lude Franke, Alexandre Reymond, and Zoltán Kutalik. The role of gene expression on human sexual dimorphism: too early to call. *bioRxiv*, April 2020.

2019 Katrin Männik, Thomas Arbogast, Maarja Lepamets, **Kaido Lepik**, Anna Pellaz, Herta Ademi, Zachary A Kupchinsky, Jacob Ellegood, Catia Attanasio, Andrea Messina, Samuel Rotman, Sandra Martin-Brevet, Estelle Dubruc, Jacqueline Chrast, Jason P Lerch, Lily R Qiu, Triin Laisk, The 16p11.2 European Consortium, The Simons VIP Consortium, The eQTLGen Consortium, R Mark Henkelman, Sébastien Jacquemont, Yann Herault, Cecilia M Lindgren, Hedi Peterson, Jean Christophe Stehle, Nicholas Katsanis, Zoltan Kutalik, Serge Nef, Bogdan Draganski, Erica E Davis, Reedik Mägi, and Alexandre Reymond. Leveraging biobank-scale rare and common variant analyses to identify ASPHD1 as the main driver of reproductive traits in the 16p11.2 locus. *bioRxiv*, July 2019.

2019 Eleonora Porcu, Sina Rüeger, **Kaido Lepik**, eQTLGen Consortium, BIOS Consortium, Federico A Santoni, Alexandre Reymond, and Zoltán Kutalik. Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat. Commun.*, 10(1):3300, July 2019.

2019 Glen James, Sulev Reisberg, **Kaido Lepik**, Nicholas Galwey, Paul Avillach, Liis Kolberg, Reedik Mägi, Tõnu Esko, Myriam Alexander, Dawn Waterworth, A Katrina Loomis, and Jaak Vilo. An exploratory phenome wide association study linking asthma and liver diseasegenetic variants to electronic health records from the Estonian Biobank. *PLoS One*, 14(4):e0215026, April 2019.

2017 **Kaido Lepik**, Tarmo Annilo, Viktorija Kukuškina, eQTLGen Consortium, Kai Kisand, Zoltán Kutalik, Pärt Peterson, and Hedi Peterson. C-reactive protein upregulates the whole blood expression of CD59 - an integrative analysis. *PLoS Comput. Biol.*, 13(9):e1005766, September 2017.

TBD Maarja Lepamets, **Kaido Lepik**, Tuuli Jürgenson, Mart Kals, Cristian Carmeli, Annique Claringbould, Murielle Bochud, Silvia Stringhini, Cisca Wijmenga, Lude Franke, Reedik Mägi, and Zoltán Kutalik. New CNV quality score enables discovering novel phenotype associations from genome-wide CNV analysis. *In preparation*.

DISSERTATIONES INFORMATICAE PREVIOUSLY PUBLISHED IN DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.
23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω-rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 lk.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
55. **Kaarel Kaljurand.** Attempto controlled English as a Semantic Web language. Tartu 2007, 162 p.
56. **Mart Anton.** Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
61. **Jüri Reimand.** Functional analysis of gene lists, networks and regulatory systems. Tartu 2010, 153 p.
62. **Ahti Peder.** Superpositional Graphs and Finding the Description of Structure by Counting Method. Tartu 2010, 87 p.
64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
66. **Mark Fišel.** Optimizing Statistical Machine Translation via Input Modification. Tartu 2011, 104 p.
67. **Margus Niitsoo**. Black-box Oracle Separation Techniques with Applications in Time-stamping. Tartu 2011, 174 p.
71. **Siim Karus.** Maintainability of XML Transformations. Tartu 2011, 142 p.
72. **Margus Treumuth.** A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. Tartu 2011, 95 p.
73. **Dmitri Lepp.** Solving simplification problems in the domain of exponents, monomials and polynomials in interactive learning environment T-algebra. Tartu 2011, 202 p.

74. **Meelis Kull.** Statistical enrichment analysis in algorithms for studying gene regulation. Tartu 2011, 151 p.
77. **Bingsheng Zhang.** Efficient cryptographic protocols for secure and private remote databases. Tartu 2011, 206 p.
78. **Reina Uba.** Merging business process models. Tartu 2011, 166 p.
79. **Uuno Puus.** Structural performance as a success factor in software development projects – Estonian experience. Tartu 2012, 106 p.
81. **Georg Singer.** Web search engines and complex information needs. Tartu 2012, 218 p.
83. **Dan Bogdanov.** Sharemind: programmable secure computations with practical applications. Tartu 2013, 191 p.
84. **Jevgeni Kabanov.** Towards a more productive Java EE ecosystem. Tartu 2013, 151 p.
87. **Margus Freudenthal.** Simpl: A toolkit for Domain-Specific Language development in enterprise information systems. Tartu, 2013, 151 p.
90. **Raivo Kolde.** Methods for re-using public gene expression data. Tartu, 2014, 121 p.
91. **Vladimir Šor.** Statistical Approach for Memory Leak Detection in Java Applications. Tartu, 2014, 155 p.
92. **Naved Ahmed.** Deriving Security Requirements from Business Process Models. Tartu, 2014, 171 p.
94. **Liina Kamm.** Privacy-preserving statistical analysis using secure multi-party computation. Tartu, 2015, 201 p.
100. **Abel Armas Cervantes.** Diagnosing Behavioral Differences between Business Process Models. Tartu, 2015, 193 p.
101. **Fredrik Milani.** On Sub-Processes, Process Variation and their Interplay: An Integrated Divide-and-Conquer Method for Modeling Business Processes with Variation. Tartu, 2015, 164 p.
102. **Huber Raul Flores Macario.** Service-Oriented and Evidence-aware Mobile Cloud Computing. Tartu, 2015, 163 p.
103. **Tauno Metsalu.** Statistical analysis of multivariate data in bioinformatics. Tartu, 2016, 197 p.
104. **Riivo Talviste.** Applying Secure Multi-party Computation in Practice. Tartu, 2016, 144 p.
108. **Siim Orasmaa.** Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience. Tartu, 2016, 186 p.
109. **Prastudy Mungkas Fauzi.** Efficient Non-interactive Zero-knowledge Protocols in the CRS Model. Tartu, 2017, 193 p.
110. **Pelle Jakovits.** Adapting Scientific Computing Algorithms to Distributed Computing Frameworks. Tartu, 2017, 168 p.
111. **Anna Leontjeva.** Using Generative Models to Combine Static and Sequential Features for Classification. Tartu, 2017, 167 p.
112. **Mozhgan Pourmoradnasseri.** Some Problems Related to Extensions of Polytopes. Tartu, 2017, 168 p.

113. **Jaak Randmets.** Programming Languages for Secure Multi-party Computation Application Development. Tartu, 2017, 172 p.
114. **Alisa Pankova.** Efficient Multiparty Computation Secure against Covert and Active Adversaries. Tartu, 2017, 316 p.
116. **Toomas Saarsen.** On the Structure and Use of Process Models and Their Interplay. Tartu, 2017, 123 p.
121. **Kristjan Korjus.** Analyzing EEG Data and Improving Data Partitioning for Machine Learning Algorithms. Tartu, 2017, 106 p.
122. **Eno Tõnisson.** Differences between Expected Answers and the Answers Offered by Computer Algebra Systems to School Mathematics Equations. Tartu, 2017, 195 p.

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

1. **Abdullah Makkeh**. Applications of Optimization in Some Complex Systems. Tartu 2018, 179 p.
2. **Riivo Kikas**. Analysis of Issue and Dependency Management in Open-Source Software Projects. Tartu 2018, 115 p.
3. **Ehsan Ebrahimi**. Post-Quantum Security in the Presence of Superposition Queries. Tartu 2018, 200 p.
4. **Ilya Verenich**. Explainable Predictive Monitoring of Temporal Measures of Business Processes. Tartu 2019, 151 p.
5. **Yauhen Yakimenka**. Failure Structures of Message-Passing Algorithms in Erasure Decoding and Compressed Sensing. Tartu 2019, 134 p.
6. **Irene Teinemaa**. Predictive and Prescriptive Monitoring of Business Process Outcomes. Tartu 2019, 196 p.
7. **Mohan Liyanage.** A Framework for Mobile Web of Things. Tartu 2019, 131 p.
8. **Toomas Krips.** Improving performance of secure real-number operations. Tartu 2019, 146 p.
9. **Vijayachitra Modhukur.** Profiling of DNA methylation patterns as biomarkers of human disease. Tartu 2019, 134 p.
10. **Elena Sügis.** Integration Methods for Heterogeneous Biological Data. Tartu 2019, 250 p.
11. **Tõnis Tasa.** Bioinformatics Approaches in Personalised Pharmacotherapy. Tartu 2019, 150 p.
12. **Sulev Reisberg.** Developing Computational Solutions for Personalized Medicine. Tartu 2019, 126 p.
13. **Huishi Yin.** Using a Kano-like Model to Facilitate Open Innovation in Requirements Engineering. Tartu 2019, 129 p.
14. **Faiz Ali Shah.** Extracting Information from App Reviews to Facilitate Software Development Activities. Tartu 2020, 149 p.
15. **Adriano Augusto**. Accurate and Efficient Discovery of Process Models from Event Logs. Tartu 2020, 194 p.
16. **Karim Baghery.** Reducing Trust and Improving Security in zk-SNARKs and Commitments. Tartu 2020, 245 p.
17. **Behzad Abdolmaleki.** On Succinct Non-Interactive Zero-Knowledge Protocols Under Weaker Trust Assumptions. Tartu 2020, 209 p.
18. **Janno Siim.** Non-Interactive Shuffle Arguments. Tartu 2020, 154 p.
19. **Ilya Kuzovkin.** Understanding Information Processing in Human Brain by Interpreting Machine Learning Models. Tartu 2020, 149 p.
20. **Orlenys López Pintado.** Collaborative Business Process Execution on the Blockchain: The Caterpillar System. Tartu 2020, 170 p.
21. **Ardi Tampuu.** Neural Networks for Analyzing Biological Data. Tartu 2020, 152 p.

22. **Madis Vasser.** Testing a Computational Theory of Brain Functioning with Virtual Reality. Tartu 2020, 106 p.
23. **Ljubov Jaanuska.** Haar Wavelet Method for Vibration Analysis of Beams and Parameter Quantification. Tartu 2021, 192 p.
24. **Arnis Parsovs.** Estonian Electronic Identity Card and its Security Challenges. Tartu 2021, 214 p.